Proceedings

THIRTY-SIXTH ANNUAL ALLERTON CONFERENCE ON COMMUNICATION, CONTROL AND COMPUTING

September 23 - 25, 1998

Allerton House, Monticello, Illinois
Sponsored by the
Coordinated Science Laboratory and the
Department of Electrical and Computer Engineering of the
University of Illinois at Urbana-Champaign

PROCEEDINGS

THIRTY-SIXTH ANNUAL ALLERTON CONFERENCE ON COMMUNICATION, CONTROL, AND COMPUTING

Tamer Başar
Bruce Hajek
Conference Co-Chairs

Conference held
September 23 - 25, 1998
Allerton House
Monticello, Illinois

Sponsored by
The Coordinated Science Laboratory
and
The Department of Electrical and Computer Engineering
of the
UNIVERSITY OF ILLINOIS
at
Urbana-Champaign

621.3811
AL5
1998

FOREWORD

These Proceedings contain papers that were presented at the Thirty-Sixth Annual Allerton Conference on Communication, Control, and Computing which was held at Allerton House, Monticello, Illinois, on September 23 through September 25, 1998. As in past years, the Conference was sponsored by the Coordinated Science Laboratory and the Department of Electrical and Computer Engineering of the University of Illinois at Urbana-Champaign.

Fifty-five invited papers and fifty-nine contributed papers were presented at the Conference in twenty-three sessions. We thank all the authors and particularly the speakers for contributing to the success of the Conference. We are also grateful to our colleagues at the University of Illinois at Urbana-Champaign: Professors T.U. Başar, R.E. Blahut, D. Jones, R. Koetter, P.R. Kumar, J. Lyding, U. Madhow, J.V. Medanić, M. Médard, D. Munson, L. Pitt, D.V. Sarwate, A. Singer, R.B. Sowers, M.W. Spong, R.S. Sreenivas, R. Srikant, A. Vardy, and P.G. Voulgaris, for organizing many of the invited sessions, reviewing the contributed papers, chairing sessions, and helping in many other ways in the smooth running of the Conference. An additional invited session was organized by Professor A.H. Sayed, from the University of California at Los Angeles; we are especially grateful to him for his efforts on behalf of the Conference. Dr. G. D. Forney, Jr. of Motorola, Inc., Professor D. Goeckel of the University of Massachusetts, Professor U. Mittra of Ohio State University, Professor G.H. Sasaki of the University of Hawaii, and Professor M. Varanasi of the University of Colorado at Boulder, chaired contributed and invited sessions at the Conference, and for this we thank them.

The plenary speaker at the Conference was Gilles Brassard of the University of Montreal, Canada, who gave an illuminating lecture titled "Quantum Information Processing." We are grateful to him for his excellent presentation. We are also indebted to the entertainment from the "Camerata IV," for giving us a wonderful evening of music.

We thank Mrs. Beverly Becker and her staff at Allerton House for cheerfully accommodating our Conference and for their constant attention to our needs. It is also a great pleasure to thank all the staff members of the Coordinated Science Laboratory who so ably handled all the practical aspects of organizing the Conference. In particular, we wish to mention Mr. Steven Williams and Ms. Susan Baxley who handled all the business affairs, and Mr. Dan Jordan who handled the sales and shipping of these Proceedings. Most of all, however, we would like to thank our Conference Secretary, Mrs. Becky Lonberger, who with the able assistance of Mrs. Francie Bridges and Mrs. Nancy Carr worked all year round in handling announcements, calls for papers, program preparation, Allerton House arrangements, registration, and the final editing of these Proceedings. The lion's share of the credit for the success of the Conference belongs to these individuals and it gives us the greatest pleasure to acknowledge this here.

We close this foreword with the countdown for the reopening of the time capsule buried Tuesday, September 30, 1997, on the occasion of the Thirty-Fifth Annual Allerton Conference. The number of years to go is 34. The capsule can be reached by taking 35 paces North from the Allerton House, along a line extending from the eastern face of the house. (It was incorrectly reported to be along a line from the western face of the house in the Proceedings of the Thirty-Fifth Conference.) The GPS coordinates (WGS84) of the time capsule are 40 degrees 00 minutes 15.39750 seconds North latitute, and 88 degrees 38 minutes 41.74234 seconds West longitude.

Tamer Başar and Bruce Hajek

Table 1: 36 YEARS OF ALLERTON CONFERENCES

Name	Dates	Chair(s)
First Annual Allerton Conference on Circuit and System Theory	November 15-17, 1963	J. B. Cruz, Jr.
Second Annual Allerton Conference on Circuit and System Theory	September 28-30, 1964	W. R. Perkins
Third Annual Allerton Conference on Circuit and System Theory	October 20-22, 1965	M. E. VanValkenburg
Fourth Annual Allerton Conference on Circuit and System Theory	October 5-7, 1966	William R. Perkins J. B. Cruz, Jr.
Fifth Annual Allerton Conference on Circuit and System Theory	October 4-6, 1967	J. B. Cruz, Jr. T. N. Trick
Sixth Annual Allerton Conference on Circuit and System Theory	October 2-4, 1968	T. N. Trick R. T. Chien
Seventh Annual Allerton Conference on Circuit and System Theory	October 8-10, 1969	A. H. Haddad G. Metze T. N. Trick
Eighth Annual Allerton Conference on Circuit and System Theory	October 7-9, 1970	A. H. Haddad B. C. Kuo G. Metze
Ninth Annual Allerton Conference on Circuit and System Theory	October 6-8, 1971	J. B. Cruz, Jr. A. H. Haddad
Tenth Annual Allerton Conference on Circuit and System Theory	October 4-6, 1972	J. B. Cruz, Jr. C. L. Chen
Eleventh Annual Allerton Conference on Circuit and System Theory	October 3-5, 1973	W. R. Perkins P. V. Kokotovic
Twelfth Annual Allerton Conference on Circuit and System Theory	October 2-4, 1974	P. V. Kokotovic E. S. Davidson
Thirteenth Annual Allerton Conference on Circuit and System Theory	October 1-3, 1975	E. S. Davidson T. N. Trick
Fourteenth Annual Allerton Conference on Circuit and System Theory	September 29-October 1, 1976	T. N. Trick M. E. VanValkenburg
Fifteenth Annual Allerton Conference on Communication, Control, and Computing	September 28-30, 1977	F. P. Preparata M. B. Pursley
Sixteenth Annual Allerton Conference on Communication, Control, and Computing	October 4-6, 1978	M. B. Pursley J. B. Cruz, Jr.
Seventeenth Annual Allerton Conference on Communication, Control, and Computing	October 10-12, 1979	J. B. Cruz, Jr. F. P. Preparata
Eighteenth Annual Allerton Conference on Communication, Control, and Computing	October 8-10, 1980	D. V. Sarwate W. R. Perkins

Nineteenth Annual Allerton Conference on Communication, Control, and Computing	September 30-October 2, 1981	W. K. Jenkins D. V. Sarwate
Twentieth Annual Allerton Conference on Communication, Control, and Computing	October 6-8, 1982	H. V. Poor W. K. Jenkins
Twenty-First Annual Allerton Conference on Communication, Control, and Computing	October 5-7, 1983	Tamer Başar Donna J. Brown H. Vincent Poor
Twenty-Second Annual Allerton Conference on Communication, Control, and Computing	October 3-5, 1984	David C. Munson, Jr. Tamer Başar
Twenty-Third Annual Allerton Conference on Communication, Control, and Computing	October 2-4, 1985	Bruce Hajek David C. Munson, Jr.
Twenty-Fourth Annual Allerton Conference on Communication, Control, and Computing	October 1-3, 1986	Michael C. Loui Juraj V. Medanic
Twenty-Fifth Annual Allerton Conference on Communication, Control, and Computing	September 30-October 2, 1987	P. R. Kumar Michael C. Loui
Twenty-Sixth Annual Allerton Conference on Communication, Control, and Computing	September 28-30, 1988	Mark W. Spong Bruce Hajek
Twenty-Seventh Annual Allerton Conference on Communication, Control, and Computing	September 27-29, 1989	Juraj V. Medanic P. R. Kumar
Twenty-Eighth Annual Allerton Conference on Communication, Control, and Computing	October 3-5, 1990	Donna J. Brown Kameshwar R. Poolla
Twenty-Ninth Annual Allerton Conference on Communication, Control, and Computing	October 2-4, 1991	George Cybenko Constantine Polychronopoulos
Thirtieth Annual Allerton Conference on Communication, Control, and Computing	September 30-October 2, 1992	Paul Van Dooren Mark W. Spong
Thirty-First Annual Allerton Conference on Communication, Control, and Computing	September 29-October 1, 1993	Dilip V. Sarwate Paul Van Dooren
Thirty-Second Annual Allerton Conference on Communication, Control, and Computing	September 28-30, 1994	W. Kenneth Jenkins Dilip V. Sarwate
Thirty-Third Annual Allerton Conference on Communication, Control, and Computing	October 4-6, 1995	Sean P. Meyn Ken Zeger
Thirty-Fourth Annual Allerton Conference on Communication, Control, and Computing	October 2-4, 1996	Sean P. Meyn W. Kenneth Jenkins
Thirty-Fifth Annual Allerton Conference on Communication, Control, and Computing	September 29-October 1, 1997	Tamer Başar Alexander Vardy
Thirty-Sixth Annual Allerton Conference on Communication, Control, and Computing	September 23-25, 1998	Tamer Başar Bruce Hajek

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i

I.A: STOCHASTIC ANALYSIS I

Organizers: B. Hajek, P.R. Kumar and R.B. Sowers
(University of Illinois at Urbana-Champaign)
Chair: S. Namachchivaya
(University of Illinois at Urbana-Champaign)

TIME AVERAGE CONTROL PROBLEMS FOR DIFFUSIONS WITH SWITCHINGS AND JUMPS 1
A. Arapostathis

AN APPROACH TO STOCHASTIC CALCULUS FOR FRACTIONAL BROWNIAN MOTION 4
T.E. Duncan

THE COMMON RANDOMNESS CAPACITY OF A FINITE NETWORK OF PROCESSORS 10
S. Venkatesan and V. Anantharam

I.B: CHANNEL CODING

Chair: D.V. Sarwate
(University of Illinois at Urbana-Champaign)

ITERATIVE MIN-SUM DECODING OF TAIL-BITING CODES 15
S. Aji, G. Horn, R.J. McEliece and M. Xu

TURBO CODES WITH SHORT FRAME LENGTHS 25
X. Lou and L.C. Pérez

AN ALGEBRAIC FRAMEWORK FOR TURBO CODES 27
B.M. Allen

TURBO CODES FOR APD-DETECTED PPM 29
J. Hamkins and M. Srinivasan

ANALYTIC TECHNIQUES FOR PERIODIC TRELLIS CODES 39
R.D. Wesel and X. Liu

AN INTERLEAVING TECHNIQUE FOR ERROR CORRECTION IN 2-D BAR CODING 49
G.F. Elmasry and Y.Q. Shi

I.C: CCOMMUNICATION NETWORKS I

Chair: G.H. Sasaki
(University of Hawaii)

FORMAL SPECIFICATION OF QUALITY OF SERVICE CONSTRAINTS FOR NETWORK PROTOCOLS 59
N. Nounou

ANALYSIS OF AN FEC/NAK APPROACH FOR SCALABLE RELIABLE MULTICAST 69
O.Z. Zeng, C.K. Mohan and C.R.P. Hartmann

STABILITY ISSUES IN HOP-BY-HOP RATE BASED CONGESTION CONTROL 79
V. Kulkarni, S. Bohacek and M.G. Safonov

TOPOLOGY AGGREGATION FOR NETWORKS WITH HIERARCHICAL STRUCTURE: A PRACTICAL APPROACH 89
Y. Shavitt

KNOWLEDGE MODELS FOR VERIFYING PROTOCOLS 97
S.L. Ricker and K. Rudie

ON FAULT-TOLERANT ROUTING IN ITERATED LINE DIGRAPH NETWORKS USING OPTICAL INTERCONNECTIONS 107
A. Sengupta

I.D: COMMUNICATIONS SYSTEMS
Chair: M. Varanasi
(University of Colorado at Boulder)

LOW-DENSITY PARITY CHECK CODES FOR FADING CHANNELS WITH MEMORY 117
A.P. Worthen and W.E. Stark

CODED MODULATION FOR PEAK POWER CONSTRAINED WIRELESS OFDM SYSTEMS 126
D.L. Goeckel

PROFILE OPTIMAL 8-QAM AND 32-QAM CONSTELLATIONS 136
X. Liu and R.D. Wesel

MULTICHANNEL RECEPTION OF DIGITAL SIGNALS OVER CORRELATED NAKAGAMI FADING CHANNELS 146
M.-S. Alouini and M.K. Simon

IMPROVED PERFORMANCE OF A PSAM SYSTEM WITH ITERATIVE FILTERING AND DECODING 156
H.-J. Su and E. Geraniotis

ML ESTIMATE BASED AVL TECHNIQUE USING SIGNAL STRENGTH MEASUREMENT WITH TIME DIVERSITY 167
I. Nomay and J. Dunham

RADIO NETWORK OPTIMIZATION BY MEANS OF NEURAL NETWORKS 177
A. Berny

II.A: STOCHASTIC ANALYSIS II
Organizers: B. Hajek, P.R. Kumar and R.B. Sowers
(University of Illinois at Urbana-Champaign)
Chair: P.R. Kumar
(University of Illinois at Urbana-Champaign)

A SHORT REVIEW OF DEGREE THEORY ON WIENER SPACE 187
A.S. Üstünel and M. Zakai

STOCHASTIC CONTROL AND LINEAR PROGRAMMING 191
L. Shepp

II.B: ITERATIVE DECODING AND TURBO CODES
Organizers: R.E. Blahut, R. Koetter and A. Vardy
(University of Illinois at Urbana-Champaign)
Chair: G.D. Forney, Jr.
(Motorola)

CODING THEOREMS FOR "TURBO-LIKE" CODES 201
D. Divsalar, H. Jin and R.J. McEliece

FINDING GOOD LDPC CODES 211
D.A. Spielman

COMPARISON OF CONSTRUCTIONS OF IRREGULAR GALLAGER CODES 220
D.J.C. MacKay, S.T. Wilson and M.C. Davey

PERFORMANCE ANALYSIS OF TURBO-EQUALIZED DICODE PARTIAL-RESPONSE CHANNEL 230
M. Öberg and P.H. Siegel

CONCATENATED CODES THAT ACHIEVE 10^{-6} FRAME ERROR RATES AT SNR'S 0.6dB FROM THE SPHERE PACKING BOUND 240
O.Y. Takeshita, O.M. Collins, P.C. Massey and D.J. Costello, Jr.

II.C.i: ADAPTIVE FILTERING

Organizer and Chair: A.H. Sayed
(University of California, Los Angeles)

STABILITY OF THE LMS ADAPTIVE FILTER BY MEANS OF A STATE EQUATION 242
V.H. Nascimento and A.H. Sayed

BLIND ADAPTIVE CROSS-POLE INTERFERENCE CANCELLATION USING FRACTIONALLY-SPACED CMA 252
W. Chung, J. Treichler and C.R. Johnson, Jr.

AN ADAPTIVE SOLUTION TO THE *EL FAROL* PROBLEM 262
W.A. Sethares and A.M. Bell

II.C.ii: CDMA COMMUNCATION NETWORKS

Chair: M. Medard
(University of Illinois at Urbana-Champaign)

OPTIMAL SEQUENCES AND SUM CAPACITY OF SYNCHRONOUS CDMA SYSTEMS 272
P. Viswanath and V. Anantharam

ADMISSIBILITY AND NETWORK CAPACITY OF POWER-CONTROLLED CDMA SYSTEMS IN FADING CHANNELS 282
J. Zhang and E.K.P. Chong

2D-RAKE COMBINER FOR DS-CDMA SYSTEMS IN THE PRESENCE OF PILOT TRANSMISSIONS 292
J. Choi and G.H. Lee

PERFORMANCE OF SCALE-TIME-CODE DIVISION MULTIPLE ACCESS FOR COMPLEX-VALUED PN SEQUENCES 294
O. Kucur and G.E. Atkin

ANALYSIS OF A DISTRIBUTED POWER CONTROL ALGORITHM 296
J.D. Herdtner and E.K.P. Chong

II.D: EQUALIZATION

Chair: D. Goeckel
(University of Massachusetts)

ADAPTIVE MULTIUSER DECISION FEEDBACK DEMODULATION FOR GSM 298
R. Ratasuk and M.L. Honig

QUASI-DETERMINISTIC SECOND ORDER BLIND SEQUENCE ESTIMATION USING ANTENNA ARRAYS 308
J. Thomas

AN INFORMATION-THEORETIC DERIVATION OF THE MMSE DECISION-FEEDBACK EQUALIZER 318
T. Guess and M.K. Varanasi

OPTIMIZED DECISION FEEDBACK EQUALIZATION FOR CONVOLUTIONAL CODING WITH REDUCED DELAY 328
J.-T. Liu and S.B. Gelfand

WHEN THE BEST DECISION-FEEDBACK EQUALIZER IS A LINEAR EQUALIZER 338
W. Shi and R.D. Wesel

PSEUDO-ORTHOGONAL BASES FOR MULTICARRIER MODULATION 340
X. Yongzhe, V.U. Reddy and K.C. Chung

MAXIMUM-LIKELIHOOD CARRIER FREQUENCY OFFSET ESTIMATION IN ORTHOGONAL FREQUENCY DIVISION MULTIPLEXING (OFDM) SYSTEMS 343
A.T. Huq, E. Panayirci and C.N. Georghiades

III.A: DECODING AND CHANNELS

Organizers: R.E. Blahut, R. Koetter and A. Vardy
(University of Illinois at Urbana-Champaign)
Chair: R.E. Blahut
(University of Illinois at Urbana-Champaign)

NEW LOWER BOUNDS ON THE DECODING ERROR PROBABILITY 353
S. Litsyn

A QUANTUM DECODING ALGORITHM OF THE SIMPLEX CODE 359
A. Barg and S. Zhou

ON TWO-DIMENSIONAL ARRAYS AND CROSSWORD PUZZLES 366
J.K. Wolf and P.H. Siegel

CODING FOR CHANNELS WITH LOCAL DATA DEPENDENT NOISE 372
R. Erstad and Ø. Ytrehus

III.B: WIRELESS COMMUNICATIONS I

Organizer: D. Jones and U. Madhow
(University of Illinois at Urbana-Champaign)
Chair: U. Madhow
(University of Illinois at Urbana-Champaign)

BLIND MULTICHANNEL EQUALIZATION FOR WIDEBAND TDMA BASED ON SUBBANDING AND THE SIMO CROSS-RELATION 381
M.D. Zoltowski and D.-F. Tseng

FURTHER RESULTS ON SPACE-TIME CODING FOR RAYLEIGH FADING 391
J. Grimm, M.P. Fitz and J.V. Krogmeier

MINIMUM-ENTROPY BLIND ACQUISITION/EQUALIZATION FOR UPLINK DS-CDMA 401
P. Schniter and C.R. Johnson, Jr.

ADAPTIVE INTERFERENCE SUPPRESSION FOR DIRECT-SEQUENCE CDMA SYSTEMS WITH LONG SPREADING CODES 411
C.D. Frank and E. Visotsky

ON THE RELATIVE ERROR PROBABILITIES OF LINEAR MULTIUSER DETECTORS 421
G.V. Moustakides and H.V. Poor

SPACE-TIME MODULATION SCHEME FOR UNKNOWN RAYLEIGH FADING ENVIRONMENTS........ 431
B.M. Hochwald and T.L. Marzetta

ITERATIVE CHANNEL ESTIMATION FOR WIDEBAND CDMA SYSTEMS.. 441
A.G. Dabak, S. Hosur and T. Schmidl

III.C: CONTROL AND OPTIMIZATION
Chair: J.V. Medanic
(University of Illinois at Urbana-Champaign)

AN INEXACT HYBRID PROJECTION-PROXIMAL AUGMENTED LAGRANGIAN ALGORITHM........... 450
C. Humes, Jr., P.J. da Silva e Silva and B.F. Svaiter

LINEAR OPTIMAL COST CUMULANT CONTROL: A k-CUMULANT PROBLEM CLASS.................... 460
K.D. Pham, S.R. Liberty and M.K. Sain

SYNTHESIS OF TIME-OPTIMAL FEEDBACK CONTROLLERS WITH MIXED CONTROL: BANG-BANG AND IMPULSE ACTIONS.. 470
J.C. Luo and E.B. Lee

A LMI APPROACH TO H_∞ CONTROL PROBLEMS FOR DESCRIPTOR SYSTEMS............................ 479
T. Miyazaki and S. Hosoe

STRATIFIED MORSE THEORY FOR STABILITY MARGIN COMPUTATION.................................... 489
N.-P. Ke

STATE FEEDBACK STABILIZATION OVER A FINITE TIME INTERVAL OF LINEAR SYSTEMS SUBJECT TO NORM BOUNDED UNCERTAINTIES.. 499
F. Amato, M. Ariola and P. Dorato

IV.A: WIRELESS COMMUNICATIONS II
Organizers: D. Jones and U. Madhow
(University of Illinois at Urbana-Champaign)
Chair: D. Jones
(University of Illinois at Urbana-Champaign)

ITERATIVE SIGNATURE ADAPTATION FOR CAPACITY MAXIMIZATION OF CDMA SYSTEMS...... 506
S. Ulukus and R.D. Yates

CAPACITY AND OPTIMAL RESOURCE ALLOCATION FOR FADING BROADCAST CHANNELS....... 516
L. Li and A. Goldsmith

ON CHANNEL MODELING FOR DELAY ANALYSIS OF PACKET COMMUNICATIONS OVER WIRELESS LINKS.. 526
M. Zorzi and R.R. Rao

CANONICAL MULTIPATH-DOPPLER COORDINATES IN WIRELESS COMMUNICATIONS................ 536
A.M. Sayeed

A NEW FRAMEWORK FOR POWER CONTROL IN WIRELESS DATA NETWORKS: GAMES, UTILITY AND PRICING.. 546
D. Famolari, N. Mandayam and D. Goodman

OPTIMUM SIGNAL DESIGN AND POWER CONTROL FOR MULTIUSER WIRELESS COMMUNICATIONS UNDER LOCATION-INVARIANT BANDWIDTH CONSTRAINTS........................ 556
M.K. Varanasi and E.A. Fain

IV.B: ALGEBRAIC AND COMBINATORIAL CODING THEORY

Organizers: R.E. Blahut, R. Koetter and A. Vardy
(University of Illinois at Urbana-Champaign)
Chair: A. Vardy
(University of Illinois at Urbana-Champaign)

ON THE CLASSIFICATION OF FORMALLY SELF-DUAL CODES 566
J.E. Fields, P. Gaborit, W.C. Huffman and V. Pless

ON THE DUAL OF MDS CONVOLUTIONAL CODES 576
J. Rosenthal and R. Smarandache

ALGEBRAIC GEOMETRY CODES ARE GENERALIZED REED-SOLOMON CODES 584
D.A. Leonard

ON THE COSET WEIGHT DISTRIBUTIONS OF Z_4-LINEAR GOETHALS CODES AND SOME RELATIONS TO KLOOSTERMAN SUMS 593
T. Helleseth and V. Zinoviev

FURTHER RESULTS ON A CONJECTURED 2-LEVEL AUTOCORRELATION SEQUENCE 598
A. Chang, T. Helleseth and P.V. Kumar

IV.C: CONTROL APPLICATIONS

Chair: W.R. Perkins
(University of Illinois at Urbana-Champaign)

CONTROLLER FRAGILITY: WHAT'S ALL THE FUSS? 600
D. Faris, T. Paré, A. Packard, K.A. Ali and J.P. How

FRICTION COMPENSATION IN MECHANICAL SYSTEMS VIA ITERATIVE LEARNING 610
S.-I. Cho, S.-J. Kim, Y.-H. Kim and I.-J. Ha

TOPOLOGY-GUIDED MOTION PLANNING 620
S. Guha, D. Suri and I. Suzuki

ACTIVE COMPENSATION OF FLEXIBILITY IN BELT-DRIVEN DRIVE TRAINS 630
R.B. Gorbet and S.A. Bortoff

IV.D.i: DETECTION AND ESTIMATION

Chair: Y. Bresler
(University of Illinois at Urbana-Champaign)

SIGNAL DETECTION IN FRACTIONAL GAUSSIAN NOISE 640
R.J. Barton and V. Nauratanmull

A DUAL SYSTEM APPROACH TO THE SOURCE LOCALIZATION PROBLEM 651
M.E. Alpay and M.H. Shor

IV.D.ii: COMMUNICATION NETWORKS II

Chair: R. Srikant
(University of Illinois at Urbana-Champaign)

A WDM RING NETWORK FOR INCREMENTAL TRAFFIC 662
G. Sasaki, O. Gerstel and R. Ramaswami

ON THE ALLOCATION OF RESOURCES OF A MULTI-MEDIA NETWORK 673
T. Heikkinen

MITIGATING THE EFFECTS OF UNCERTAINTY IN JOIN-THE-SHORTEST-QUEUE RESOURCE ALLOCATION SCHEMES THROUGH PRIORITIZED DISPERSAL 683
Y. Birk and N. Bloch

A FRAMEWORK FOR ADAPTIVE SERVICE GUARANTEES 693
R. Agrawal, R.L. Cruz, C.M. Okino and R. Rajan

IV.E: COMPLEXITY AND CRYPTOGRAPHY
Chair: L. Pitt
(University of Illinois at Urbana-Champaign)

TIME-BOUNDED PREFIX KOLMOGOROV COMPLEXITY 703
D.W. Juedes and J.H. Lutz

MULTISIGNATURE SCHEMES FOR VARIOUS GROUP STRUCTURES 713
H. Doi, E. Okamoto and M. Mambo

PLENARY SESSION

QUANTUM INFORMATION PROCESSING 723
G. Brassard

V.A: CONTROL ISSUES IN LOCOMOTION
Organizer and Chair: M.W. Spong
(University of Illinois at Urbana-Champaign)

PLANNING AND CONTROL FOR PLANAR BATTING AND HOPPING 724
C.K. Black and K.M. Lynch

STABILIZATION OF BALANCE IN 3-D BIPEDAL LOCOMOTION 734
A.D. Kuo and C.E. Bauby

EXPLOITING NATURAL DYNAMICS IN THE CONTROL OF A PLANAR BIPEDAL WALKING ROBOT 739
J.E. Pratt and G.A. Pratt

MUSCULOSKELETAL STIFFNESS IN LOCOMOTION 749
C.T. Farley

METABOLIC ENERGY CONSUMPTION DURING ANIMAL LOCOMOTION 757
R. Kram

V.B: MULTIUSER DETECTION
Chair: U. Mitra
(Ohio State University)

ARCHITECTURAL IMPLEMENTATION ISSUES IN A WIDEBAND RECEIVER USING MULTIUSER DETECTION 765
N. Zhang, C. Teuscher, H. Lee and B. Brodersen

ITERATIVE TECHNIQUES FOR DS/CDMA MULTIPATH CHANNEL ESTIMATION 772
E. Ertin, U. Mitra and S. Siwamogsatham

USING LINEAR PREDICTION IN JOINT BLIND MULTIUSER DETECTION AND BLIND CHANNEL EQUALIZATION FOR CDMA SYSTEMS 782
H.H. Fan and X. Li

ON THE DIMENSIONAL LIMITATIONS OF LINEAR MULTIUSER DETECTION 790
S. Gollamudi and Y.-F. Huang

V.C: QUANTIZATION AND SOURCE/CHANNEL CODING
Chair: A. Singer
(University of Illinois at Urbana-Champaign)

JOINT SOURCE-CHANNEL CODING OF A GAUSSIAN MIXTURE SOURCE OVER THE GAUSSIAN BROADCAST CHANNEL 800
Z. Reznic, R. Zamir and M. Feder

SEPARATION OF RANDOM NUMBER GENERATION AND RESOLVABILITY 810
K. Visweswariah, S. Kulkarni and S. Verdú

AN ASYMPTOTIC SOLUTION TO THE SUPPORT REGION OF A MINIMUM MEAN SQUARE-ERROR GAUSSIAN QUANTIZER.. 820
S. Na and D.L. Neuhoff

Time Average Control Problems for Diffusions with Switchings and Jumps

Ari Arapostathis

Dept. of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78712-1084
ari@mail.utexas.edu

Abstract

We study the long-term average optimal control problem for switching diffusions with jumps. We characterize the optimal policy as a minimizing selector of the Hamiltonian associated with the Hamilton-Jacobi-Bellman equations.

1 Summary

We study a stochastic control system with a discrete parameter [5], i.e., a system with state $\bigl(X(t), S(t)\bigr) \in \mathbb{R}^d \times \mathcal{S}$, where $\mathcal{S} = \{1, 2, \dots, N\}$, and control parameter $u(t)$, for which the continuous state $X(t)$ evolves according to a jump diffusion process

$$dX(t) = f\bigl(X(t), S(t), u(t)\bigr)dt + \sigma\bigl(X(t), S(t)\bigr)dW(t) + \int_{\mathbb{R}^m\setminus\{0\}} g\bigl(X(t), S(t), \xi\bigr)\tilde{\nu}(dt, d\xi), \quad (1)$$

where $\bigl\{W(t) : t \geq 0\bigr\}$ is a standard Wiener process in $\mathbb{R}^d$ and $\bigl\{\tilde{\nu}(t, \cdot) : t \geq 0\bigr\}$, is a martingale measure in $\mathbb{R}^m \setminus \{0\}$ independent of $\bigl\{W(t) : t \geq 0\bigr\}$. On the other hand, the discrete state $S(t)$ behaves as a controlled Markov chain with a transition matrix $\Lambda = [\lambda_{ij}]$, i.e., roughly speaking,

$$P\bigl(S(t+\delta t) = j \mid S(t) = i\bigr) = \lambda_{ij}\bigl(X(t), u(t)\bigr)\delta t + o(\delta t), \quad i \neq j\,. \quad (2)$$

A precise mathematical model for this system can be obtained as in [2]. Let U be the control space, which is assumed compact, and $\mathcal{V}$ denote the space of probability measures on its Borel σ-field, $\mathfrak{B}(U)$, endowed with the Prohorov topology. Also, for a Polish space Y, let $\mathfrak{M}(Y)$ denote the set of nonnegative integer-valued σ-finite measures on $\mathfrak{B}(Y)$. The model involves a complete probability space $(\Omega, \mathfrak{F}, P)$ with the following objects:

(i) $W(\cdot) = \bigl[W_1(\cdot), \dots, W_d(\cdot)\bigr]'$ is a d-dimensional standard Wiener process.

(ii) $\mathfrak{p}(dt, dz)$ is an $\mathfrak{M}(\mathbb{R}_+ \times \mathbb{R})$-valued Poisson random measure with intensity $dt \times m(dz)$, where m is the Lebesgue measure on $\mathbb{R}$.

(iii) $\tilde{\nu}(t, d\xi)$ is a martingale measure in $\mathbb{R}^m \setminus \{0\}$, corresponding to a standard Poisson random measure $\nu(t, d\xi)$, with associated Lévy measure π.

(iv) $\mathfrak{p}(\cdot,\cdot)$, $W(\cdot)$, and $\tilde{\nu}(\cdot,\cdot)$ are independent.

(v) $v(\cdot)$ is an *admissible (control) policy*, that is, a $\mathcal{V}$-valued process and nonanticipative, i.e., the σ-algebras

$$\mathcal{F}_t^v = \sigma\{v(s), s \le t\},$$
$$\mathcal{F}_{[t,\infty)}^{W,\mathfrak{p},\nu} = \sigma\big\{W(s) - W(t), \mathfrak{p}(A,\cdot), \nu(A',\cdot) : A, A' \in \mathfrak{B}([s,\infty)),\ s \ge t\big\},$$

are independent, for each $t \in \mathbb{R}$.

The switching jump diffusion process is governed by the Itô equation:

$$dX(t) = f\big(X(t), S(t), v(t)\big)\,dt + \sigma\big(X(t), S(t)\big)\,dW(t) + \int_{\mathbb{R}^m\setminus\{0\}} g\big(X(t), S(t), \xi\big)\tilde{\nu}(dt, d\xi) \tag{3}$$
$$dS(t) = \int_{\mathbb{R}} h\big(X(t), S(t-), v(t), z\big)\,\mathfrak{p}(dt, dz)\,,$$

where h is a suitably defined integer-valued function [2]. We wish to minimize a.s. over all admissible policies, the functional

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T c\big(X(t), S(t), v(t)\big)\,dt\,. \tag{4}$$

We make the following assumptions:

Assumption 1. The following hold:

(a) $f(x,k,u)$, $\sigma_{ij}(x,k)$, $\lambda_{ij}(x,k)$, $g(x,k,\xi)$ and the cost function $c(x,k,u)$ are continuous and Lipschitz in x, uniformly in the rest of the parameters. Also, $g(0,\cdot,\cdot)$ is bounded.

(b) $\sigma_{ij}(\cdot,\cdot)$ is uniformly elliptic.

(c) The matrix $\Lambda(x,u)$ is irreducible for all $(x,u) \in \mathbb{R}^d \times U$.

Assumption 2. The Lévy measure π is finite and if we define

$$M_{x,k}(B) := \pi\big(\xi : g(x,k,\xi) \in B\big)\,, \quad B \in \mathfrak{B}(\mathbb{R}^n)$$

then $M_{x,k}(B)$ is absolutely continuous with respect to the Lebesgue measure, uniformly on compacta.

For $\varphi \in W^{2,p}_{\ell oc}(\mathbb{R}^d \times \mathcal{S})$, $v \in \mathcal{V}$, we write

$$L^v\varphi(x,k) = L_k^v\varphi(x,k) + \sum_{j=1}^N \lambda_{kj}(x,v)\varphi(x,j), \tag{5}$$

where $L_k^v = \mathcal{A}_k^v + \mathcal{B}_k^v$, and

$$\mathcal{A}_k^v = \frac{1}{2}\sum_{i,j,\ell=1}^d \sigma_{i\ell}(x,k)\sigma_{j\ell}(x,k)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{j=1}^d f_j(x,k,v)\frac{\partial}{\partial x_j}\,,$$
$$\mathcal{B}_k^v\varphi(x,k) = \int_{R^n\setminus\{0\}} \big[\varphi(x+z,k) - \varphi(x,k) - z\cdot\nabla\varphi(x,k)\big]\,M_{x,k}(dz)\,.$$

Under Assumptions 1 and 2, the method in [6] can be used to establish that, if v is a Markov policy, then (3) has an a.s. unique strong solution, such that $\big(X(t), S(t)\big)$ is a strong Feller process with differential generator L^v, given in (5).

For $x \in \mathbb{R}^d$, $i \in \mathcal{S}$, let $V_\alpha(x,i)$ be the discounted value function with discount factor $\alpha > 0$, i.e.,

$$V_\alpha(x,i) = \inf_{v\in\Pi} E^v_{x,i}\Big[\int_0^\infty e^{-\alpha t} c\big(X(t), S(t), v(t)\big)\,dt\Big].$$

We establish the Hamilton-Jacobi-Bellman (HJB) equations, under the following *almost monotone* assumption:

$$\liminf_{\|x\|\to\infty} \Big[\inf_{u\in U} c(x,k,u) - \limsup_{\alpha\to 0} \alpha V_\alpha(x,k)\Big] > 0\,, \quad \forall k \in \mathcal{S}\,. \tag{6}$$

The problem for diffusions without jumps, i.e., $g \equiv 0$, was studied in [3]. In this paper, Harnack's inequality for the associated elliptic system [1] was crucial in obtaining the HJB equations via the vanishing discount method. The ergodic control of a reflected diffusion with jumps, is treated in [4]. For the problem at hand we will make use of the following condition.

Assumption 3. For each $k \in \mathcal{S}$, exists an increasing sequence $\{\mathcal{W}^k_n\}_{n=1}^\infty$ of compact subsets of $\mathbb{R}^m$, which cover $\mathbb{R}^m$, and such that the support of $M_{x,k}$ lies in $\mathcal{W}^k_n$ for all $x \in \mathcal{W}^k_n$, $n \in \mathbb{N}$.

Under Assumptions 1–3 and (6), using the vanishing discount approach, we obtain a pair $(V,\rho) \in W^{2,p}_{\ell oc}(\mathbb{R}^d \times \mathcal{S}) \times \mathbb{R}$, $2 \le p < \infty$, satisfying the HJB equation

$$\inf_{u\in U}\big\{L^u V(x,i) + c(x,i,u)\big\} = \rho\,. \tag{7}$$

References

[1] A. Arapostathis, M. K. Ghosh and S. I. Marcus, "Harnack's inequality for cooperative, weakly coupled elliptic systems," submitted to *Communications on Partial Differential Equations.*

[2] M. K. Ghosh, A. Arapostathis and S. I. Marcus, "Optimal control of switching diffusions with application to flexible manufacturing systems," *SIAM J. Control and Optimization,* vol. 31, pp. 1183–1204, 1993.

[3] M. K. Ghosh, A. Arapostathis and S. I. Marcus, "Ergodic control of switching diffusions," *SIAM J. Control and Optimization,* vol. 35, pp. 1952–1988, November 1997.

[4] J. L. Menaldi and M. Robin, "An Ergodic control problem for reflected diffusion with jump," *IMA Journal of Mathematical Control and Information,* vol. 1, pp. 309-322, 1984.

[5] A. V. Skorohod, *Asymptotic Methods in the Theory of Stochastic Differential Equations,* American Mathematical Society, Providence, 1989.

[6] A. Yu. Veretennikov, "On strong solutions of Itô stochastic equations with jumps," *Theory of Probability and its Applications,* vol. 32, pp. 148–152, 1988.

An Approach to Stochastic Calculus for Fractional Brownian Motion*

T. E. Duncan
Mathematics Department
University of Kansas
Lawrence, KS 66045
duncan@math.ukans.edu

Abstract

Fractional Browinian motion is a family of Gaussian processes that has played an increasingly important role in various applications. A description of some results for a stochastic calculus for fractional Brownian motion is given here. These results are proved in [2]. A stochastic calculus is important for the effective use of fractional Brownian motion in applications.

1 Introduction

Fractional Brownian motion has a relatively long history in probability. In applications it was developed by Hurst [5], [6]. Let $0 < H < 1$. It is well-known that there is a Gaussian stochastic process $(B_t^H, t \geq 0)$ such that

$$\mathbb{E}\left(B_t^H\right) = 0, \quad \mathbb{E}(B_t^H B_s^H) = \frac{1}{2}\left\{|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right\} \tag{1.1}$$

for all $s, t \in \mathbb{R}_+$. This process is called the fractional Brownian motion with Hurst parameter H. To simplify the presentation, it is always assumed that the fractional Brownian motion is 0 at $t = 0$.

If $H = 1/2$, then the corresponding fractional Brownian motion is the usual standard Brownian motion. If $H > 1/2$, then the process $(B_t^H, t \geq 0)$ exhibits a long-range dependence, that is, if $r(n) = \text{cov}(B_1^H, B_{n+1}^H - B_n^H)$, then $\sum_{n=1}^{\infty} r(n) = \infty$. A fractional Brownian motion is also self-similar, that is, $B_{\alpha t}^H$ has the same probability law as $\alpha^H B_t^H$. A process satisfying this property is called a self-similar process with the Hurst parameter H.

To use fractional Brownian motion effectively in applications it is important to have a stochastic calculus for these processes. A stochastic integral with respect to fractional Brownian motion which has expectation zero and is different from some other definitions (e.g., [1], [7]) is given here. A change of variables or Itô formula is given. The definition of stochastic integral uses the Wick product. These results as well as some others are proved in [2].

---

*Research supported partially by NSF Grant DMS 9623439

2 Fractional Brownian Motion

Let $\Omega = C_0(\mathbb{R}_+, \mathbb{R})$ be the space of continuous functions ω from $\mathbb{R}_+$ to $\mathbb{R}$ with $\omega(0) = 0$. There is a probability measure P^H on $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is the Borel σ-algebra such that on the probability space $(\Omega, \mathcal{F}, P^H)$ the coordinate process $B^H : \Omega \to \mathbb{R}$ defined as

$$B_t^H(\omega) = \omega(t), \quad \omega \in \Omega$$

is a Gaussian process satisfying (1.1). The process $(B_t^H, t \geq 0)$ is called the canonical fractional Brownian motion with Hurst parameter H. In this paper only the canonical process and its associated probability space are used. Throughout this paper it is assumed that $H \in (\frac{1}{2}, 1)$ is arbitrary but fixed. Clearly if $H = 1/2$, the fractional Brownian motion is the standard Brownian motion.

Let $\phi(s,t) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ be given by

$$\phi(s,t) = H(2H-1)|s-t|^{2H-2}. \tag{2.1}$$

Many results of this paper can be extended to a more general $\phi(s,t)$ that is symmetric and positive definite, so ϕ in (2.1) is given as a function of two variables and not their difference. Let f be a Borel measurable (deterministic) function from $\mathbb{R}_+$ to $\mathbb{R}$. Then $f \in L_\phi^2(\mathbb{R}_+)$ if

$$|f|_\phi^2 := \int_0^\infty \int_0^\infty f(s)f(t)\phi(s,t)dsdt < \infty. \tag{2.2}$$

The stochastic integral with respect to fractional Brownian motions for deterministic kernels is simple.

Lemma 2.1 *If $f, g \in L_\phi^2(\mathbb{R}_+)$, then $\int_0^\infty f_s dB_s^H$ and $\int_0^\infty g_s dB_s^H$ are well-defined zero mean, Gaussian random variables with variances $|f|_\phi^2$ and $|g|_\phi^2$ respectively and*

$$\mathbb{E}\left(\int_0^\infty f_s dB_s^H \int_0^\infty g_s dB_t^H\right) = \int_0^\infty \int_0^\infty f(s)g(t)\phi(s,t)dsdt = \langle f, g\rangle_\phi. \tag{2.3}$$

The lemma is verified in [3]. It can be proved directly by verifying it for simple functions $\sum_{i=1}^n a_i \chi_{[t_i,t_{i+1}]}(s)$ and then proceeding with a passage to the limit.

3 Stochastic Integration for Fractional Brownian Motions

Let $(\Omega, \mathcal{F}, P^H)$ be the probability space from Section 2 where a fractional Brownian motion with Hurst parameter H is well-defined. The probability measure P^H depends on H. Throughout this paper the Hurst parameter H is fixed such that $1/2 < H < 1$. Since H is fixed, the probability measure is denoted by P.

Let $L^p(\Omega, \mathcal{F}, P) = L^p$ be the space of all random variables $F : \Omega \to \mathbb{R}$ such that

$$\|F\|_p := (\mathbb{E}|F|^p)^{1/p} < \infty.$$

Let $L_\phi^2(\mathbb{R}_+) = \left\{f \mid f : \mathbb{R}_+ \to \mathbb{R},\ |f|_\phi^2 := \int_0^\infty \int_0^\infty f_s f_t \phi(s,t)dsdt < \infty\right\}$. Often for notational simplicity $L_\phi^2(\mathbb{R}_+)$ is denoted by L_ϕ^2. For any $f \in L_\phi^2$, define $\varepsilon : L_\phi^2 \to L^1(\Omega, \mathcal{F}, P)$ as

$$\varepsilon(f) := \exp\left\{\int_0^\infty f_t dB_t^H - \frac{1}{2}\int_0^\infty\int_0^\infty f_s f_t \phi(s,t) ds dt\right\}. \tag{3.1}$$

If $f \in L^2_\phi$, then $\varepsilon(f) \in L^p(\Omega, \mathcal{F}, P)$ for each $p \geq 1$ and is called an exponential function. The Hilbert space L^2_ϕ is naturally associated with the Gaussian process, fractional Brownian motion, by the approach using an abstract Wiener space. The Hilbert space plays a basic role for questions of absolute continuity and the exponential function (3.1) is a Radon-Nikodym derivative for a translate of the fractional Brownian motion.

Let $\mathcal{E}$ be the linear span of the exponentials, that is,

$$\mathcal{E} = \left\{\sum_{k=1}^{n} a_k \varepsilon(f_k), \quad n \in \mathbb{N},\ a_k \in \mathbb{R},\ f_k \in L^2_\phi(\mathbb{R}_+) \text{ for } k \in \{1, \cdots, n\}\right\}. \tag{3.2}$$

Theorem 3.1 *$\mathcal{E}$ is a dense set of $L^p(\Omega, \mathcal{F}, P)$ for each $p \geq 1$. In particular, $\mathcal{E}$ is a dense set of $L^2(\Omega, \mathcal{F}, P)$.*

For a functional F in $L^p(\Omega, \mathcal{F}, P)$ $(p \geq 1)$ and a function $g \in L^2_\phi$,

$$F(B^H_\cdot + \int_0^\cdot \int_0^\infty \phi(u,v) g_v du dv)$$

is well-defined. An analogue of the Malliavin derivative [8] is introduced.

Definition 3.1 *The ϕ-derivative of F in the direction of Φg is defined as*

$$D_{\Phi g} F(\omega) = \lim_{\delta \to 0} \frac{1}{\delta}\left\{F(\omega + \delta \int_0^\cdot \int_0^\infty \phi(u,v) g(v) dv du) - F(\omega)\right\} \tag{3.3}$$

if the limit exists in $L^p(\Omega, \mathcal{F}, P)$. Furthermore if there is a process f_s such that

$$D_{\Phi g} F = \int_0^\infty f_s g_s ds \quad a.s.$$

for all $g \in L^2_\phi$ then F is said to be ϕ-differentiable, $D^\phi F$ is said to exist and f_s is denoted by $D^\phi_s F$, that is,

$$D_{\Phi g} F(\omega) = \int_0^\infty D^\phi_s F(\omega) g_s ds\,.$$

The higher derivatives can be defined in a similar manner.

Definition 3.2 *Let $F : [0,T] \times \Omega \to \mathbb{R}$ be a stochastic process. The process F is said to be ϕ-differentiable if for each $t \in [0,T]$, $F(t, \cdot)$ is ϕ- differentiable and $D^\phi_s F_t$ is jointly measurable.*

It is easy to see that if $f : \mathbb{R} \to \mathbb{R}$ is a smooth function and $F : \Omega \to \mathbb{R}$ is ϕ-differentiable then $f(F)$ is also ϕ-differentiable and

$$D_{\Phi g} f(F) = f'(F) D_{\Phi g} F \tag{3.4}$$

and

$$D_{\Phi g_1} D_{\Phi g_2} f(F) = f'(F) D_{\Phi g_1} D_{\Phi g_2} F + f''(F) D_{\Phi g_1} F D_{\Phi g_2} F\,.$$

The following rules for differentiation are useful later

$$D_{\Phi g} \int_0^\infty f_s dB_s^H = \int_0^\infty \int_0^\infty \phi(u,v) f_u g_v du dv = \langle f, g \rangle_\phi; \tag{3.5}$$

$$D_s^\phi \int_0^\infty f_u dB_u^H = \int_0^\infty \phi(u,s) f_u du = (\Phi f)(s); \tag{3.6}$$

$$D_{\Phi g}\varepsilon(f) = \varepsilon(f) \int_0^\infty \int_0^\infty \phi(u,v) f_u g_v du dv = \varepsilon(f)\langle f, g\rangle_\phi; \tag{3.7}$$

$$D_s^\phi \varepsilon(f) = \varepsilon(f) \int_0^\infty \phi(u,s) f_u du = \varepsilon(f)(\Phi f)(s), \tag{3.8}$$

where $f,\ g \in L_\phi^2$. Now the Wick product $\diamond$ of two functionals is introduced. To extend the theory of stochastic calculus for Brownian motions to the fractional Brownian motions, the Wick calculus for Gaussian processes (or Gaussian measures) is used. The Wick product of two exponentials $\varepsilon(f)$ and $\varepsilon(g)$ is defined as

$$\varepsilon(f) \diamond \varepsilon(g) = \varepsilon(f+g). \tag{3.9}$$

Since for distinct $f_1,\ f_2,\ \ldots,\ f_n$ in L_ϕ^2, $\varepsilon(f_1), \varepsilon(f_2), \ldots, \varepsilon(f_n)$ are linearly independent, this definition can be extended to define the Wick product $F \diamond G$ of two functionals F and G in $\mathcal{E}$.

Consider an arbitrary partition of $[0,T]$, $\pi : 0 = t_0 < t_1 < t_1 < \cdots < t_n = T$. First the following Riemann sum is given using the Wick product introduced above

$$S_\pi = \sum_{i=0}^{n-1} F_{t_i} \diamond (B_{t_{i+1}}^H - B_{t_i}^H).$$

From (3.9), it easily follows that that for any F and G in $\mathcal{E}$, $\mathbb{E}(F \diamond G) = \mathbb{E}(F)\mathbb{E}(G)$. This identity extends to more general F and G such that $F \diamond G$ is well-defined (e.g., [4], p. 83). Thus for any partition π,

$$\begin{aligned} \mathbb{E}\left(\sum_{i=0}^{n-1} F_{t_i}^\pi \diamond (B_{t_{i+1}}^H - B_{t_i}^H)\right) &= \sum_{i=0}^{n-1} \mathbb{E}\left(F_{t_i}^\pi \diamond (B_{t_{i+1}}^H - B_{t_i}^H)\right) \\ &= \sum_{i=0}^{n-1} \mathbb{E}\left(F_{t_i}^\pi\right) \mathbb{E}\left(B_{t_{i+1}}^H - B_{t_i}^H\right) = 0 \end{aligned}$$

Let $\mathcal{L}(0,T)$ be the family of stochastic processes on $[0,T]$ such that $F \in \mathcal{L}(0,T)$ if $\mathbb{E}|F|_\phi^2 < \infty$, F is ϕ-differentiable, the trace of $\left(D_s^\phi F_t,\ 0 \le s \le T,\ 0 \le t \le T\right)$ exists and $\mathbb{E}\int_0^T (D_s^\phi F_s)^2 ds < \infty$ and for each sequence of partitions $(\pi_n,\ n \in \mathbb{N})$ such that $|\pi_n| \to 0$ as $n \to \infty$

$$\sum_{i=0}^{n-1} \mathbb{E}\left\{\int_{t_i^{(n)}}^{t_{i+1}^{(n)}} |D_s^\phi F_{t_i} - D_s^\phi F_s| ds\right\}^2$$

and

$$\mathbb{E}|F^\pi - F|_\phi^2$$

tend to 0 as $n \to \infty$, where $\pi_n : 0 = t_0^{(n)} < t_1^{(n)} < \cdots < t_{n-1}^{(n)} < t_n^{(n)} = T$.

The following result summarizes the above construction of a stochastic integral.

Theorem 3.2 *Let $F:\ \mathbb{R}_+ \times \Omega \to \mathbb{R}$ be a stochastic process such that $F \in \mathcal{L}(0,T)$. The limit*

$$\lim_{|\pi| \to 0} \sum_{i=0}^{n-1} F_{t_i}^\pi \diamond (B_{t_{i+1}}^H - B_{t_i}^H)$$

exists and this limit is defined as $\int_0^T F_s \delta B_s^H$. Moreover, $\mathbb{E}\int_0^T F_s \delta B_s^H = 0$ and

$$\mathbb{E}|\int_0^T F_s \delta B_s^H|^2 = \mathbb{E}\left\{\left(\int_0^T D_s^\phi F_s ds\right)^2 + \int_0^T \int_0^T F_u F_v \phi(u,v) du dv\right\}.$$

An analogue of the stochastic integral of Stratonovich type $\int_0^t F_s dB_s^H$ is also introduced. This type of integral is related to the integrals introduced by Lin [7] and Dai and Heyde [1].

Definition 3.3 *Let* $(\pi_n,\, n \in \mathbb{N})$ *be a sequence of partitions of* $[0,t]$ *such that* $|\pi_n| \to 0$. *If* $\sum_{i=0}^{n-1} f(t_i^{(n)})(B^H(t_{i+1}^{(n)}) - B^H(t_i^{(n)}))$ *converges in* $L^2(\Omega, \mathcal{F}, P)$ *to the same limit for all such* $(\pi_n, n \in \mathbb{N})$, *then this limit is called the stochastic integral of Stratonovich type and the limit is denoted by* $\int_0^t f(s)dB(s)$.

Theorem 3.3 *If* $F \in \mathcal{L}(0,T)$, *then the stochastic integral of Stratonovich type* $\int_0^t F_s dB_s^H$ *exists and the following equality is satisfied*

$$\int_0^t F_s dB_s^H = \int_0^t F_s \delta B_s^H + \int_0^t D_s^\phi F_s ds \quad a.s.$$

Now an analogue of the Itô formula is established, that is, a chain rule for the integral introduced in the last section. Let $f : \mathbb{R} \to \mathbb{R}$ be a twice continuously differentiable function with bounded second derivative. Then for a partition $\{t_0, t_1, \ldots, t_n\}$ of $[0,T]$ it follows by Taylor's formula that

$$\begin{aligned} f(B_T^H) - f(0) &= \sum_{i=0}^{n-1} \left[f(B_{t_{i+1}}^H) - f(B_{t_i}^H) \right] \\ &= \sum_{i=0}^{n-1} f'(B_{t_i}^H) \left[B_{t_{i+1}}^H - B_{t_i}^H \right] + \frac{1}{2} \sum_{i=0}^{n-1} f''(\xi_i) \left[B_{t_{i+1}}^H - B_{t_i}^H \right]^2 \\ &= \sum_{i=0}^{n-1} f'(B_{t_i}^H) \diamond \left[B_{t_{i+1}}^H - B_{t_i}^H \right] + \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} D_s^\phi f'(B_{t_i}^H) ds \\ &\quad + \frac{1}{2} \sum_{i=0}^{n-1} f''(\xi_i) \left[B_{t_{i+1}}^H - B_{t_i}^H \right]^2 \\ &= I_1 + I_2 + I_3, \end{aligned}$$

where $\xi_i \in (B_{t_i}^H, B_{t_{i+1}}^H)$. Since is is assumed that $H > 1/2$, it follows that $I_3 \to 0$ in $L^2(\Omega, \mathcal{F}, P)$. By the definition of the stochastic integral introduced in the preceding section, the first term I_1 converges to $\int_0^T f'(B_s^H) \delta B_s^H$ in L^2. By the chain rule for the ϕ-differentiation operator, it follows that for $s \in [t_i, t_{i+1})$

$$\begin{aligned} D_s^\phi f'(B_{t_i}^H) &= f''(B_{t_i}^H) D_s^\phi B_{t_i}^H \\ &= f''(B_{t_i}^H) \int_0^{t_i} \phi(u,s) du \\ &= H f''(B_{t_i}^H) \left[s^{2H-1} - (s - t_i)^{2H-1} \right] \end{aligned}$$

Thus the second sum in the three sums from Taylor's formula converges to

$$H \int_0^T s^{2H-1} f''(B_s^H) ds$$

in L^2. The following chain rule formula is obtained.

Theorem 3.4 *If $f : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable function with bounded derivatives, then*

$$f(B_T^H) - f(B_0^H) = \int_0^T f'(B_s^H)\delta B_s^H + H\int_0^T s^{2H-1} f''(B_s^H)ds \quad a.s. \tag{3.10}$$

It is interesting to note that this formula implies the usual Itô formula for Brownian motion when $H = 1/2$ is formally substituted in (3.10).

References

[1] W. Dai and C. C. Heyde, "Itô formula with respect to fractional Brownian motion and its application," *J. Appl. Math. Stoch. Anal.*, vol. 9, pp. 439–448, 1996.

[2] T. E. Duncan, Y. Hu, and B. Pasik-Duncan, Stochastic calculus for fractional Brownian motion. I: Theory, preprint.

[3] G. Gripenberg and I. Norros, "On the prediction of fractional Brownian motion," *J. Appl. Prob.*, vol. 33, pp. 400–410, 1996.

[4] H. Holden, B. Øksendal, J. Ubøe, and T. S. Zhang, *Stochastic Partial Differential Equations, a Modeling, White Noise Functional Analysis*, Birkhäuser, 1996.

[5] H. E. Hurst, "Long-term storage capacity in reservoirs," *Trans. Amer. Soc . Civil Eng.*, vol. 116, pp. 400–410, 1951.

[6] H. E. Hurst, "Methods of using long-term storage in reservoirs," *Proc. Inst. Civil Engineers Part I*, Chapter 5, pp. 519–590, 1956.

[7] S. J. Lin, "Stochastic analysis of fractional Brownian motions," *Stochastics and Stochastic Rep.*, vol. 55, pp. 121–140, 1995.

[8] S. Watanabe, *Stochastic Differential Equation and Malliavin Calculus*, Tata Institute of Fundamental Research, Springer, 1984.

The Common Randomness Capacity of a Finite Network of Processors *

Sivarama Venkatesan and Venkat Anantharam

Dept. of EECS
University of California, Berkeley
Berkeley, CA94720-1772
ananth@eecs.berkeley.edu

Abstract

A finite number of processors interconnected by an arbitrary network of independent, point-to-point, discrete memoryless channels wish to generate common randomness by interactive, synchronized communication. We determine the common randomness capacity of such a network in a Shannon theoretic sense, i.e. the maximum number of bits of randomness that all the processors can agree on, per step of communication.

1 Introduction

Two or more communicating agents are said to have *common randomness* if there is a random variable, e.g., a random bit string, whose value is known to all of them, with probability extremely close to 1. This notion of shared randomness turns out to be of significance in many problems of information theory. For example, common randomness available to a transmitter and receiver allows them to use random codes for data transmission, which can outperform deterministic codes in certain situations, e.g., with arbitrarily varying channels [1], [6]. In the theory of identification over noisy channels [4], [3], [5], the maximum achievable identification rate is essentially determined by the amount of common randomness that the transmitter and receiver can set up. Common randomness can also significantly reduce the communication complexity of certain distributed computations [7], [9]. Finally, *secret* common randomness available to a transmitter and receiver allows them to communicate securely over a channel with eavesdroppers [2], [8].

An important observation is the possibility of exploiting *channel noise* to generate common randomness. To see how channel noise could actually be useful in this context, consider a situation where there are two agents A and B connected to each other in both directions by a pair of channels. As an extreme case, assume that neither agent has access to any external sources of randomness (such as a random bit generator). Even then, A and B may be able to generate common randomness! The intuition is this: suppose A transmits some known input sequence to B. If the channel from A to B is noisy, then the resulting output sequence seen by B will be random. Since the input sequence is

---

*Research supported by the National Science Foundation under grants NCR 94-22513 and IRI 95-05561.

known, B could somehow "cancel" out its effect on the output sequence, and extract the randomness due to noise. Now, if the channel from B to A has positive Shannon capacity, then B could reliably convey the randomness thus obtained to A, using suitable encoding techniques; A and B would then have *common* randomness.

A natural question that arises now is: what is the maximum *rate*, in bits per step of communication, at which the two agents can generate common randomness this way from the noise on the two channels, i.e., what is the common randomness *capacity* of the given pair of channels? Here we would require the probability of not sharing a common random value to vanish asymptotically in the number of steps of communication. This question was posed and answered in [11], under the assumption that the two channels are independently operating discrete memoryless channels (DMC's). The main result of [11] is that A and B can generate

$$\max_{X_A, X_B} \quad \{\min\left[H(Y_B|X_B), I(X_A;Y_A)\right] + \min\left[H(Y_A|X_A), I(X_B;Y_B)\right]\} \tag{1}$$

bits of common randomness per step of communication over such a pair of channels. In (1), (X_A, Y_A) and (X_B, Y_B) are random input-output pairs for the A-to-B and B-to-A DMC's, respectively, and the maximum is over all possible distributions for the inputs X_A and X_B. The rate in (1) was shown to be optimal with a "strong" converse, which was proved by developing a novel typical sequence machinery for interactive communication. Further, it was shown that the common randomness capacity in the presence of independent, discrete memoryless sources of external randomness at the two terminals could also be derived from (1).

In the special case where both channels are binary symmetric with crossover probabilities p and q, respectively, the expression in (1) reduces to $\min\{h(p)+h(q), 2-h(p)-h(q)\}$, where $h(\cdot)$ is the binary entropy function. Observe that this is 0 if and only if either $h(p) = h(q) = 0$ (no randomness on either channel) or $h(p) = h(q) = 1$ (plenty of randomness but no ability to agree); moreover the capacity equals its maximum of 1 bit per step when $h(p) = h(q) = 1/2$. For details and other examples, see [11].

In this invited talk we consider a network of agents, represented by a digraph $G = (V, E)$. The vertex set V of this digraph is just the set of agents, and the edge set $E \subseteq V \times V$ describes their interconnections — $(u, v) \in E$ means there is a channel whose input is controlled by u, and whose output is seen by v. As in [11], we assume that these channels are all DMC's, and that they operate independently. Communication occurs simultaneously on all the channels, and in synchronism (i.e., there is a common clock).

As in the two-agent case, there are basically two steps here in the process of generating common randomness. The first step is for the agents to bring channel noise into play by communicating over the channels, so that each agent can then extract randomness from the observed channel outputs. The second step is for each agent to convey reliably the randomness thus obtained to each of the other agents, using suitable encoding techniques. However, important new features show up in the general problem that are not present in the problem for two agents. In the two-agent case, there is only one path along which an agent can deliver randomness to the other, and this path consists of a single channel. Also, the flows of randomness originating from the two agents do not interact — each is confined to a different channel. In contrast, in an arbitrary network of channels, there could be several paths from one agent to each of the others, many of these consisting of more than one channel, and all these paths could be used simultaneously to deliver randomness. Moreover, a given channel could be on several different paths, which means that the flows of randomness from different agents must interact. For these reasons, the problem of optimally disseminating randomness from each agent to all the others is quite non-trivial in the general situation.

Nevertheless, the final result we prove has an intuitively appealing form. For simplicity, let us assume that the processors do not have access to any external sources of randomness; the results are easily generalized to allow for such sources. Consider the digraph $G = (V, E)$. A *path* from vertex u to vertex w is a set of $k \geq 1$ edges

$$\{(v_0, v_1), (v_1, v_2), (v_2, v_3), \ldots, (v_{k-1}, v_k)\},$$

with $v_0 = u$ and $v_k = w$. A *circuit* is a path from some vertex to itself. Note that a self-loop (v, v) constitutes a circuit by itself. An *arborescence rooted at* v is a set T of edges, with the following properties: (i) no edge in T enters v; (ii) for each $u \neq v$, exactly one edge in T enters u; and (iii) T does not contain any circuits. It can be verified easily that the above properties imply that (iv) $|T| = |V| - 1$; (v) for each $u \neq v$, there is a *unique* path in T from v to u; and (vi) the edges of T form a spanning tree in the undirected graph underlying G. We will denote the set of all arborescences rooted at v by $\mathcal{T}(v)$, and let $\mathcal{T} = \bigcup_{v \in V} \mathcal{T}(v)$.

If $\mathbf{Q} = (Q(y|x) : (x, y) \in \mathcal{X} \times \mathcal{Y})$ is the matrix of transition probabilities of a discrete memoryless channel (DMC) with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, P is a probability distribution on $\mathcal{X}$, and X and Y are random variables with the joint distribution $Pr\{X = x, Y = y\} = P(x)Q(y|x)$, let us write $H(\mathbf{Q}|P)$ and $I(P; \mathbf{Q})$ for $H(Y|X)$ and $I(X; Y)$, respectively. To each edge $e \in E$ there is associated a DMC, whose input alphabet is denoted $\mathcal{X}_e$, output alphabet is denoted $\mathcal{Y}_e$, and transition probabilities are denoted $\mathbf{Q}_e = (Q_e(y|x) : (x, y) \in \mathcal{X}_e \times \mathcal{Y}_e)$.

For each $e \in E$, let P_e be a probability distribution on the input alphabet $\mathcal{X}_e$ of channel e, and let $\mathbf{P} = (P_e : e \in E)$. Let the vectors $\mathbf{a}(\mathbf{P}) \in \mathbf{R}_+^V$ and $\mathbf{b}(\mathbf{P}) \in \mathbf{R}_+^E$ be given by

$$a_v(\mathbf{P}) \stackrel{\text{def}}{=} \sum_{e \in \delta_{in}(v)} H(\mathbf{Q}_e | P_e), \tag{2}$$

$$b_e(\mathbf{P}) \stackrel{\text{def}}{=} I(P_e; \mathbf{Q}_e), \tag{3}$$

and let $\mathcal{R}(\mathbf{P})$ be the polyhedron of all vectors $\mathbf{r} \in \mathbf{R}_+^{\mathcal{T}}$ satisfying the following constraints:

$$\sum_{T \in \mathcal{T}(v)} r_T \leq a_v(\mathbf{P}), \quad \text{for each } v \in V; \tag{4}$$

$$\sum_{T : e \in T} r_T \leq b_e(\mathbf{P}), \quad \text{for each } e \in E. \tag{5}$$

Note that the constraint in (5) can be ignored if e is a self-loop, because the LHS is then a summation over an empty set (no arborescence contains a self-loop), and is therefore equal to zero.

Let

$$C(\mathbf{P}) \stackrel{\text{def}}{=} \max_{\mathbf{r} \in \mathcal{R}(\mathbf{P})} \sum_{T \in \mathcal{T}} r_T, \tag{6}$$

$$C \stackrel{\text{def}}{=} \max_{\mathbf{P}} C_*(\mathbf{P}). \tag{7}$$

Our main result is the the common randomness capacity of the network of processors is precisely C. The proof involves a direct (achievability) part and a converse part.

For the converse we are again able to prove what is commonly known as a "strong converse" in Shannon theory. The proof of the converse requires an initial use of linear programming duality to identify a finite set of information theoretic inequalities that need to be established.

A formal statement of the main results was presented at the 1997 Asilomar Conference on on Signals, Systems and Computers [13], where a formal definition of "common randomness" is given and the proof is sketched in outline. Full proofs are available in [12]. A journal paper on this work was prepared subsequently, and this is currently under review for publication. The work was part of the doctoral dissertation of the first author at Cornell University [14], and the major portion of it was done while he was an exchange scholar at the University of California at Berkeley.

References

[1] R. Ahlswede. "Elimination of Correlation in Random Codes for Arbitrarily Varying Channels", *Z. Wahrsch. Verw. Gebiete*, Vol. 33, 1978, pp. 159-175.

[2] R. Ahlswede and I. Csisz"à"r. "Common Randomness in Information Theory and Cryptography, Part 1: Secret Sharing", *IEEE-IT*, Vol. 39, No. 4, July 1993, pp. 1121 -1132.

[3] R. Ahlswede and G. Dueck. "Identification via Channels", *IEEE-IT*, Vol. 35, No. 1, January 1989, pp. 15 -29.

[4] R. Ahlswede and G. Dueck. "Identification in the Presence of Feedback - A Discovery of New Capacity Formulas", *IEEE-IT*, Vol. 35, No. 1, January 1989, pp. 30 -36.

[5] R. Ahlswede and B. Verboven. "On Identification via Multiway Channels with Feedback", *IEEE-IT*, Vol. 37, No. 6, November 1991, pp. 1519 -1526.

[6] I. Csiszar and P. Narayan. "The Capacity of the Arbitrarily Varying Channel Revisited: Positivity Constraints", *IEEE-IT*, Vol. 34, No. 2, March 1988, pp. 181-193.

[7] L. Lovász. "Communication Complexity: A Survey", In *Paths, Flows and VLSI layout*, B.H. Korte et. al, Editors, Springer-Verlag, 1990.

[8] U.M. Maurer". "Secret Key Agreement by Public Discussion from Common Information", *IEEE-IT*, Vol. 39, No. 3, May 1993, pp. 733 -742.

[9] A. Orlitsky and A. El Gamal. "Communication Complexity", In *Complexity in Information Theory*, Y. Abu-Mostafa, Editor, Springer-Verlag, 1988.

[10] A. Orlitsky and A. El Gamal, "Average and Randomized Communication Complexity", *IEEE-IT*, Vol. 36, No. 1,January 1990, pp. 3 -16.

[11] S. Venkatesan and V. Anantharam. "The Common Randomness Capacity of a Pair of Independent Discrete Memoryless Channels", *Electronics Research Laboratory*, University of California, Berkeley, UCB/ERL M95/85, September 1995.

[12] S. Venkatesan and V. Anantharam. "Generating Common Randomness in an Arbitrary Network of Channels : Capacity Formulas and some Combinatorial Results",

Electronics Research Laboratory, University of California, Berkeley, UCB/ERL M97/9, January 1997.

[13] S. Venkatesan and V. Anantharam, “The Common Randomness Capacity of a Finite Network of Channels”, *Proceedings of the 31st Asilomar Conference on Signals, Systems and Computers*, November 2 -5 1997, Asilomar, California (Invited), pp. 1359 -1363.

[14] S. Venkatesan, “Generating Common Randomness in an Arbitrary Network of Channels : Capacity Formulas and Combinatorial Results”, *Doctoral Dissertation*, School of Electrical Engineering, Cornell University, Ithaca, New York, January 1998.

Iterative Min-Sum Decoding of Tail-biting Codes[1]

Srinivas Aji, Gavin Horn, Robert McEliece, and Meina Xu
Department of Electrical Engineering,
California Institute of Technology,
Pasadena, California 91125, USA
e-mail: {mas, gavinh, rjm, mxu}@systems.caltech.edu

Abstract: By invoking a form of the Perron-Frobenius theorem for the "min-sum" semiring, we obtain a union bound on the performance of iterative decoding of tail-biting codes for both the AWGN channel and the BSC. This bound shows that iterative decoding will be nearly optimum, at least for high SNRs, if and only if the minimum "pseudodistance" of the code is at least the ordinary minimum distance. We prove that the minimum pseudodistance of any 2-segment "pseudocodeword" must be at least the ordinary minimum distance. Unfortunately, the minimum pseudodistance of a 3(or more)-segment pseudocodeword can be less than the ordinary minimum distance. We will present a tail-biting trellis for the $(24, 12, 8)$ Golay which has a pseudoweight less than the minimum distance of the code.

1 Introduction

Because of the remarkable success of the iterative turbo-decoding algorithm [7], many coding researchers have been focusing on the study of other sub-optimal iterative decoding algorithms. Perhaps the simplest such algorithm is the iterative decoding of tail-biting codes. In this paper we show that iterative min-sum decoding of a tail-biting code will be effective if and only if the minimum "pseudoweight" of the code is at least its ordinary minimum weight. Closely related results were discovered independently by Wiberg in his thesis [15] and by Forney et al, in [10].

This paper is an extension of [2]. Besides containing some new results on pseudoweights, this paper also includes results for the BSC as well.

2 Perron-Frobenius Theorem for the min-sum semiring

In this section we will state without proof a "Perron-Frobenius" theorem for the min-sum semiring, which explains the behavior of the iterative decoding algorithm to be presented in section 4. (Cf. the usual "sum-product" P.-F. theorem, e.g. [11, Theorem 4.5.12]).

Let A be an $s \times s$ irreducible matrix with entries from $R \cup \{\infty\}$, with rows and columns indexed by $\{1, 2, \ldots, s\}$, and let G be the corresponding weighted digraph. Assume that among all simple closed paths in G, there is a unique one with minimum average edge weight, and that this "critical cycle" is in fact a self-loop of weight ρ at vertex 1. (We summarize this condition by saying that A has a "simple eigenvalue.") Then for n sufficiently large, with min-sum arithmetic,

$$A^n = \rho^n E.$$

---

[1] This work was partially supported by NSF grant no. NCR-9505975, AFOSR grant no. 5F49620-97-1-0313, and a grant from Qualcomm, Inc. Gavin Horn's contribution was also supported by an NSERC and a Fessenden scholarship.

Here E is a fixed $s \times s$ "rank one" matrix, i.e., $E_{i,j} = x_i y_j$, where $\mathbf{x} = (x_1, \ldots, x_s)$ and $\mathbf{y} = (y_1, \ldots, y_s)$ are right and left "eigenvectors" for A with corresponding "eigenvalue" ρ.

The numbers x_i and y_j have the following geometric interpretation. If we reduce each entry in A by ρ, and if we denote this modified matrix by A' and the corresponding weighted digraph by G', then x_i is the least weight of any path in G' from vertex i to vertex 1, and y_j is the least weight of any path from vertex 1 to vertex j.

3 Tail-Biting Codes and Pseudo-Codewords

In this section we give a brief introduction to tail-biting codes. For further details, we refer the reader to [8].

A tail-biting trellis is a finite, labeled, digraph in which the vertices are partitioned into n classes, $\Sigma_0, \ldots, \Sigma_{n-1}$, each class being indexed by an element of $Z_n = \{0, 1, \ldots, n-1\}$, the cyclic group of order n. (All index arithmetic is done modulo n.) If E is an edge, we denote the initial vertex of E by $\mathrm{init}(E)$ and the final vertex of E by $\mathrm{fin}(E)$. An edge E must have $\mathrm{init}(E) \in \Sigma_k$ and $\mathrm{fin}(E) \in \Sigma_{k+1}$, for some $k \in Z_n$. The label of such an edge, denoted $\mathrm{out}(E)$ (for "output"), belongs to a finite alphabet A_k. If P is a path, the label of P, denoted $\mathrm{out}(P)$, is the concatenation of the labels of the edges comprising P. We call $\mathrm{out}(P)$ the output of the path P.

An *L-segment tail-biting path* P is a trellis path of length Ln for which $\mathrm{init}(P) = \mathrm{fin}(P)$. The *code* generated by the tail-biting trellis is the set of outputs of the the one-segment tail-biting paths. A *pseudocodeword* is the output of any tail-biting path of 1 or more segments.

4 Iterative Decoding of Tail-biting Codes

The iterative min-sum decoding algorithm for tail-biting codes is discussed explicitly in [13, 5, 9, 14]. Our view is that it is an application of the Generalized Distributive Law [4, 1], as applied to a junction graph with just one cycle [3].

In any case, if $\mathbf{y}$ is the received noisy codeword, after a finite number of iterations, the decoder will "lock on" to the pseudocodeword nearest to $\mathbf{y}$, which is called the *dominant pseudocodeword* in [9]. This follows from the min-sum Perron-Frobenius theorem (alternatively see [14] or [9]). Here the appropriate matrix A has entry $a_{i,j}$ given by $a_{i,j} = \min\{p(\mathbf{y}|out(P) : \mathrm{init}(P) = i, \mathrm{fin}(P) = j\}$. A ML decoder will compute $\min_i\{a_{i,i}\}$, since that corresponds to the most likely tail-biting codeword. On the other hand, a two-way iterative min-sum decoding algorithm will converge after a finite number of iterations, to the same result, *provided A has a simple eigenvalue.* In coding terms, this condition amounts to saying that there is a unique nearest pseudocodeword to $\mathbf{y}$, which is in fact a codeword. This fact allows us to bound the probability of decoder error, using the familiar union bound argument.

5 The Union Bound for Iterative Decoding on the AWGN Channel

In this section we restrict attention to binary linear (tail-biting) codes, being used with BPSK modulation on an additive white Gaussian channel. We will use the insights gained in the previous section (the decoder converges to the nearest pseudocodeword) to obtain a union bound on the decoder word error probability.

Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$ be an L-segment $(0,1)$ pseudocodeword, $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_L)$ be the transmitted signal of $\mathbf{x}$ where 0 is mapped into 0 and 1 is mapped into $2\sqrt{E_s}$. Assume the all zero codeword $\mathbf{0}$ is transmitted, and $\mathbf{r}$ is received. Then a decoder error occurs, i.e., the decoder prefers $\mathbf{x}$ to $\mathbf{0}$ when $\frac{1}{L}\sum_{i=1}^{L} |\mathbf{r} - \mathbf{s}_i|^2 \leq |\mathbf{r}|^2$, or equivalently, when $2(\sum_{i=1}^{L} \mathbf{s}_i)^T \mathbf{r} > \sum_{i=1}^{L} |\mathbf{s}_i|^2$. Hence the probability of decoding error is

$$Q\left(\frac{\sum_{i=1}^{L} |\mathbf{s}_i|^2}{2|\sum_{i=1}^{L} \mathbf{s}_i|} / \sigma\right), \tag{1}$$

where $Q(t) = (1/\sqrt{2\pi}) \int_t^\infty e^{-s^2/2} ds$, $\sigma^2 = N_0/2$. If c_j (the jth column sum) is defined to be $c_j = \sum_{i=1}^{L} x_{i,j}$ for $j = 1, \ldots, n$, then $\sum_{i=1}^{L} |\mathbf{s}_i|^2 = (\sum_j c_j) 4E_s$, $|\sum_{i=1}^{L} \mathbf{s}_i| = \sqrt{(\sum_j c_j^2) 4E_s}$, and eq. (1) becomes

$$Q\left(\sqrt{\frac{(\sum_j c_j)^2}{\sum_j c_j^2} 2E_s/N_0}\right). \tag{2}$$

In view of (2), we define the pseudoweight $w(\mathbf{x})$ to be

$$w(\mathbf{x}) = \frac{(\sum_j c_j)^2}{\sum_j c_j^2}.$$

Thus for example the three-segment pseudocodeword (0000) (0101) (0011) has $c_1 = 0$, $c_2 = c_3 = 1$, and $c_4 = 2$, so that its pseudoweight is $(0+1+1+2)^2/(0^2+1^2+1^2+2^2) = 8/3$. Note that the pseudoweight of an ordinary (1-segment) codeword is the same as its weight as usually defined.

Let C denote the set of all codewords, and P denote the set of all *simple* pseudocodewords (a simple pseudocodeword is one whose underlying path does not pass through the same vertex twice). The above argument implies a union bound on the (iterative min-sum) decoder word error probability P_E^{IT}:

$$P_E^{\mathrm{IT}} \leq \sum_{\mathbf{x} \in P} Q(\sqrt{2Rw(\mathbf{x})E_b/N_0}), \tag{3}$$

$$P_E^{\mathrm{ML}} \leq \sum_{\mathbf{x} \in C} Q(\sqrt{2Rw(\mathbf{x})E_b/N_0}). \tag{4}$$

(Here for completeness we have included the ordinary union bound on P_E^{ML}, the maximum-likelihood word error probability.)

In general it is not easy to compute the pseudoweight-enumerator for a given code. However, in the next section we will do so for the $(8,4,4)$ Hamming code.

6 The $(8,4,4)$ Hamming Code on AWGN Channel

In [8, section 5.2], an optimal tail-biting trellis for the extended $(8,4,4)$ binary Hamming code is constructed, with state-complexity profile $(2,4,4,4,2,4,4,4)$. We have used this trellis to experiment with the iterative min-sum decoding algorithm.

In Figure 1, we have plotted the actual performance (bit error probability) of an ML decoder and an iterative min-sum decoder (five iterations) for the $(8,4,4)$ Hamming code for values of E_b/N_0 ranging from 0 dB to 9 dB in increments of 0.5 dB. We see no measurable difference in performance, although we know that theoretically the iterative algorithm is not as good as ML because of the presence of pseudocodewords.

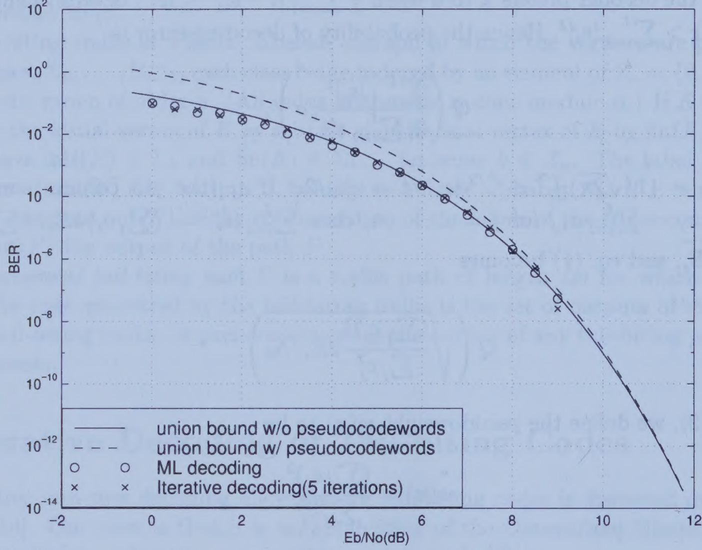


Figure 1: (8,4,4) tail-biting Hamming code union bound on AWGN channel

The pseudoweight enumerator for the $(8,4,4)$ Hamming code, as represented by the minimal tail-biting trellis from [8] is given in the table below. In the first row is the ordinary weight enumerator, i.e., a list of the weights of the codewords (the one-segment pseudocodewords). In the second row is the pseudoweight enumerator for the 64 two-segment pseudocodewords. (There can be no simple pseudocodewords with more than two segments, because the trellis has state complexity 2 at two indices.)

pseudoweight	0	1	2	3	4	$4\frac{1}{2}$	5	6	$6\frac{1}{4}$	7	$7\frac{2}{17}$	8
W.E	1				14							1
P.W.E						32			30		2	

In Figure 1, we have plotted the bounds in eqs. (4) and (3), using the data from the table, modified to give bounds on bit error probability. These bounds are asymptotically equal. This is because the leading term in both bounds is the same, because the minimum pseudoweight (in this case 4.5) is strictly larger than the minimum weight (in this case 4).

7 Pseudoweights on the AWGN Channel

Because of the union bound in eq. (3), if the minimum pseudoweight is at least the minimum weight for a given code, then the performance of iterative min-sum decoding will be asymptotically the same as ML decoding. The problem remains to find the pseudoweights. The following lemma shows that any 2-segment pseudocodeword has pseudoweight at least the minimum weight when the 2 segments are not identical.

Lemma 1 *Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be n dimensional vectors in Euclidian space. Then*

$$\frac{|\mathbf{x}_1|^2 + |\mathbf{x}_2|^2}{|\mathbf{x}_1 + \mathbf{x}_2|} \geq |\mathbf{x}_1 - \mathbf{x}_2|.$$

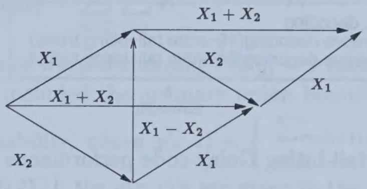


Figure 2:

Proof:

$$\begin{aligned} 2|\mathbf{x}_1 - \mathbf{x}_2||\mathbf{x}_1 + \mathbf{x}_2| &\leq |\mathbf{x}_1 - \mathbf{x}_2|^2 + |\mathbf{x}_1 + \mathbf{x}_2|^2 \\ &= 2|\mathbf{x}_1|^2 + 2|\mathbf{x}_2|^2 \end{aligned}$$

(see Figure 2 for the last equality). □

Theorem 1 *The pseudoweight of any 2-segment nonzero pseudocodeword must be at least the minimum weight if the 2 segments are not identical.*

Proof: From the definition of pseudoweight, a 2-segment $w(\mathbf{x})$ can be expressed as: $\sqrt{w(\mathbf{x})} = \frac{|\mathbf{x}_1|^2+|\mathbf{x}_2|^2}{|\mathbf{x}_1+\mathbf{x}_2|} \geq |\mathbf{x}_1 - \mathbf{x}_2|$ (the last inequality is by the Lemma). But $\mathbf{x}_1 - \mathbf{x}_2$ is an ordinary (tail-biting) codeword, and $|\mathbf{x}_1 - \mathbf{x}_2|^2$ is the codeword weight. □

If we assume the tail-biting trellis is "n-observable", i.e., that P can be inferred from out(P) for path of length $\geq n$, then even if $\mathbf{x}_1 = \mathbf{x}_2$, the conclusion of Theorem 1 holds.

For tail-biting trellises where the minimum number of states is greater than 2, the 2-segment pseudocodewords are not all the pseudocodewords. However, the pseudoweight of a 3(or more)-segment pseudocodeword can be less than the minimum distance. Köetter and Vardy have shown an example in which the pseudoweight of a 3-segment pseudocodeword is less than the minimum distance of the code[12]. Anderson directed us to a 64-state tail-biting trellis for the extended (24, 12, 8) Golay code [6], and we have found that this tail-biting trellis has a 4-segment pseudoweight of weight 7.36, which is less than the minimum distance 8 of the code. In [8], an optimal 16-state tail-biting trellis is constructed for the same Golay code. (For this 16-state tail-biting trellis, we have found that the 3-segment and 4-segment pseudocodewords are at least the minimum distance.)

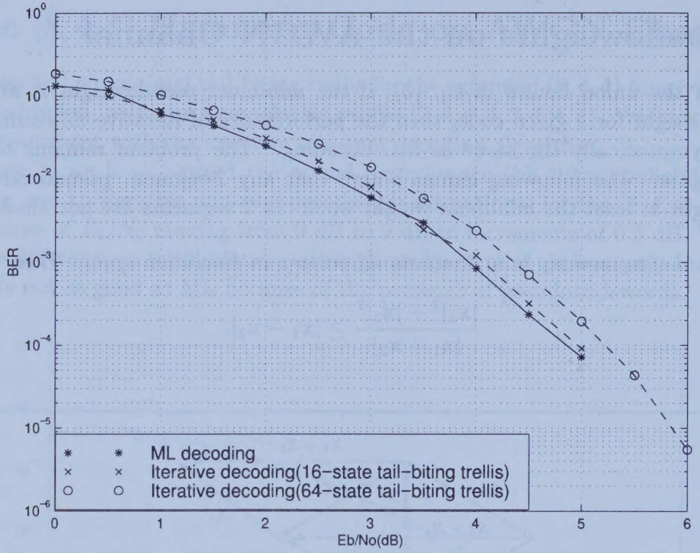


Figure 3: (24,12,8) tail-biting Golay code performance on AWGN channel

We have plotted the min-sum iterative decoding performance of both trellises as well as the ML decoding performance in Figure 3. While there is not much difference between the performance of the ML decoding and the min-sum iterative decoding of the 16-state tail-biting trellis, there is about 0.5 dB difference between the ML decoding and the min-sum iterative decoding of the 64-state tail-biting trellis, due to the low weight pseudocodeword.

8 The Union Bound for Iterative Decoding on the BSC

With $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$ and $\mathbf{c} = (c_1, \ldots, c_n)$ as defined in Section 5, we will obtain a union bound on the decoder word error probability for BSCs.

Let the all zero codeword $\mathbf{0}$ be transmitted, and let $\mathbf{y} = (y_1, \ldots, y_n)$ be the received version of $\mathbf{x}$. Then for a general two-input memoryless channel, a decoding error occurs if

$$\begin{aligned} \prod_{i=1}^{L} P(\mathbf{y}|x_i) &\geq P(\mathbf{y}|\mathbf{0})^L \\ \prod_{i=1}^{n} P(y_i|1)^{c_i} &\geq \prod_{i=1}^{n} P(y_i|0)^{c_i} \\ \left(\sum_{i:y_i=1} c_i - \sum_{i:y_i=0} c_i \right) \log \frac{P(y_i|1)}{P(y_i|0)} &\geq 0. \end{aligned} \tag{5}$$

On a BSC with crossover probability $p(p < 1-p)$, eq. (5) becomes

$$\left(\sum_{i:y_i=1} c_i - \sum_{i:y_i=0} c_i\right)\log\frac{1-p}{p} \geq 0$$

$$\sum_{i:y_i=1} c_i - \sum_{i:y_i=0} c_i \geq 0. \tag{6}$$

Let C denote the set of all codewords, and $\mathcal{C}$ denote the set of all $\mathbf{c}$ of the set of all simple pseudocodewords P. Then iterative min-sum decoder word error probability $P_{E_{BSC}}^{\mathrm{IT}}$ is:

$$P_{E_{BSC}}^{\mathrm{IT}} \leq \sum_{\mathbf{c}\in\mathcal{C}}\sum_{i=1}^{n} q(\mathbf{c},i)p^i(1-p)^{n-i}, \tag{7}$$

$$P_{E_{BSC}}^{\mathrm{ML}} \leq \sum_{\mathbf{c}\in C}\sum_{i=1}^{n} q(\mathbf{c},i)p^i(1-p)^{n-i}, \tag{8}$$

where $q(\mathbf{c},i)$ is the number of ways of choosing i of $y_j = 1$ such that eq. (6) is true. (Here for completeness we have included the ordinary union bound on $P_{E_{BSC}}^{\mathrm{IT}}$, the maximum-likelihood word error probability, where $q(\mathbf{c},i) = \begin{cases} \sum_{r=(i+1)/2}^{i}\binom{i}{r} & \text{if } i \text{ is odd} \\ q(\mathbf{c},i-1) & \text{if } i \text{ is even.}\end{cases}$) Thus for example, for $\mathbf{c} = (12121211)$, the $q(\mathbf{c},i)$'s are given in the table below.

i	$q(\mathbf{c},i)$
1	0
2	0
3	$\binom{3}{3}$
4	$\binom{3}{3}\binom{5}{1}+\binom{3}{2}\binom{5}{2}$
5	$\binom{3}{3}\binom{5}{2}+\binom{3}{2}\binom{5}{3}+\binom{3}{1}\binom{5}{4}$
6	$\binom{3}{3}\binom{5}{3}+\binom{3}{2}\binom{5}{4}+\binom{3}{1}\binom{5}{5}$
7	$\binom{3}{3}\binom{5}{4}+\binom{3}{2}\binom{5}{5}$
8	$\binom{3}{3}\binom{5}{5}$

Note that the iterative min-sum decoder word error probability union bound on BSC $P_{E_{BSC}}^{IT}$ is the same as $P_{E_{BSC}}^{ML}$ for ordinary codewords where $c_j \in \{0,1\}$, and $q(\mathbf{c},i) = \binom{n}{i}$. For pseudocodewords, c_j is a measure of the weight of bit j such that an evil channel may choose to flip the bit j which has large c_j.

Note that the min-sum iterative decoding union bound for the AWGN channel eq. (3) can also be derived from eq. (5) with the AWGN channel conditional probabilities $P(y|0) = \frac{1}{\sqrt{N_o\pi}}\exp\frac{-(y-1)^2}{N_o}$, and $P(y|1) = \frac{1}{\sqrt{N_o\pi}}\exp\frac{-(y+1)^2}{N_o}$.

9 The $(8,4,4)$ Hamming Code on the BSC

For the same $(8,4,4)$ Hamming code tail-biting trellis as in Section 6, there are 64 pseudocodewords, and the $\mathbf{c}$'s are in one of the following three forms or a permutation of them: $\mathbf{c}_1 = (12121211)$, $\mathbf{c}_2 = (01011210)$, and $\mathbf{c}_3 = (01211212)$. The corresponding $q(\mathbf{c}_j,i)$'s are given in the table below.

i	$q(\mathbf{c}_1, i)$	$q(\mathbf{c}_2, i)$	$q(\mathbf{c}_3, i)$
1	0	0	0
2	0	5	0
3	1	10	13
4	35	5	34
5	55	56	21
6	28	28	28
7	8	8	8
8	1	1	1

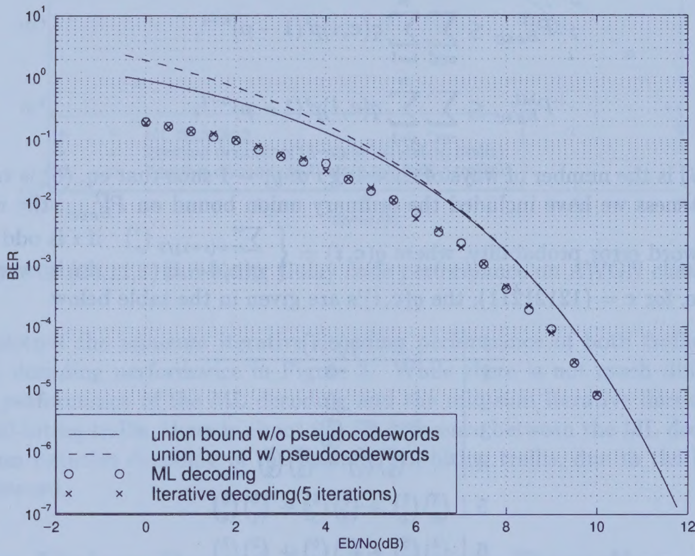


Figure 4: (8,4,4) tail-biting Hamming code union bound on BSC

In Figure 4, we have plotted the bounds in eqs.(7) and (8), using the data from the table, modified to give bounds on bit error probability. This bounds are asymptotically equal. This is because the leading term in both bounds is the same, because the minimum pseudoweight is the same as the minimum weight of the code(in both casees 4).

In Figure 4, we have plotted the actual performance (bit error probability) of an ML decoder and an iterative min-sum decoder (five iterations) for the $(8, 4, 4)$ Hamming code for values of E_b/N_0 ranging from 0 dB to 10 dB in increments of 0.5 dB. We see no measurable difference in performance.

10 Pseudoweights on the BSC

Define $\hat{\imath}$ as the minimum of i such that $q(\mathbf{c}, i) \neq 0$ over all i and $\mathbf{c}$, and the "minimum BSC pseudoweight" as $2\hat{\imath}$. If the minimum BSC pseudoweight is at least as large as the minimum Hamming weight of the code, then the iterative min-sum asymptotic performance is not degraded. We have the following theorem for pseudoweight of any 2-segment pseudocodewords.

Theorem 2 *The minimum BSC pseudoweight of any nonzero 2-segment pseudocodeword must be at least the minimum weight if the 2 segments are not identical.*

Proof: WLOG, rearrange the pseudocodeword fragments $\mathbf{x}_1$ and $\mathbf{x}_2$ so that

$$\begin{aligned}
\mathbf{x}_1 &= 1\ldots1\ 0\ldots0\ 1\ldots1\ 0\ldots0 \\
\mathbf{x}_2 &= 1\ldots1\ 1\ldots1\ 0\ldots0\ 0\ldots0 \\
\mathbf{x}_1 \oplus \mathbf{x}_2 &= \underbrace{0\ldots0}_{n_1}\ \underbrace{1\ldots1\ 1\ldots1}_{n_2}\ 0\ldots0 \\
\mathbf{c} &= \underbrace{2\ldots2}_{n_1}\ \underbrace{1\ldots1\ 1\ldots1}_{n_2}\ 0\ldots0,
\end{aligned}$$

where $\mathbf{c}$ is the column sum(ordinary sum) of the pseudocodeword fragments. Let d_{min} denote the ordinary minimum distance, and $\hat{\imath}$ as defined in the beginning of the section. Since $\mathbf{x}_1 \oplus \mathbf{x}_2$ is an ordinary codeword, $n_2 \geq d_{min}$. Hence, it is suffice to show $\hat{\imath} \geq \lceil \frac{n_2}{2} \rceil$. There are two cases to be considered.

Case 1: $n_1 \geq \lceil \frac{n_2}{2} \rceil$.

$$\begin{aligned}
2\hat{\imath} &\geq \frac{1}{2}(2n_1 + n_2) \\
&\geq \lceil \frac{n_2}{2} \rceil + \frac{n_2}{2}.
\end{aligned}$$

Thus, $\hat{\imath} \geq \lceil \frac{n_2}{2} \rceil$.

Case 2: $n_1 < \lceil \frac{n_2}{2} \rceil$. Let integer k such that $k + n_1 = \hat{\imath}$. Then,

$$\begin{aligned}
2n_1 + k &\geq \frac{1}{2}(2n_1 + n_2) \\
n_1 + k &\geq \frac{n_2}{2}.
\end{aligned}$$

Thus, $\hat{\imath} \geq \lceil \frac{n_2}{2} \rceil$. □

Again, the pseudoweight of an 3(or more)-segment pseudocodeword can be less than the minimum distance. In particular, the two tail-biting trellises in section 7 have pseudoweight 6, which is less than the minimum distance 8 of the Golay code.

11 Conclusions

The results in this paper strongly suggest that the excellent experimental performance reported for iterative min-sum decoding of tail-biting codes is due to the fact that the minimum pseudoweight of most tail-biting codes is strictly larger than the ordinary minimum weight. It remains a challenging problem to produce a practical algorithm for computing the minimum pseudoweight. In any case, because of the union bound, we can say with confidence that if, for a given code, the minimum pseudoweight is indeed greater than the minimum weight, then the performance of iterative min-sum decoding will be asymptotically the same as ML decoding.

References

[1] S. M. Aji, *Graphical Models and Iterative Decoding.* PH.D. thesis, Caltech, 1998.

[2] S. M. Aji, G. B. Horn, R. J. McEliece, and M. Xu, "Iterative min-sum decoding of tail-biting codes," Proc. IEEE Information Theory Workshop, Killarney Ireland, June 1998. pp. 68-69.

[3] S. M. Aji, G. B. Horn, and R. J. McEliece, "On the convergence of iterative decoding on graphs with a single cycle," Proc. CISS 1998(Princeton, N.J., March 1998).

[4] S. M. Aji and R. J. McEliece, "The generalized distributive law," preliminary versions presented at ISIT97 and ISCTA97. Current version available at `www.systems.caltech.edu/EE/Faculty/rjm/`, submitted to *IEEE Trans. Inform. Theory.*

[5] J. B. Anderson and S. M. Hladik, "Tail-biting MAP decoders," *IEEE J. Select. Areas Comm.*, vol. 16, no. 2 (Feb. 1998).

[6] J. B. Anderson, personal communication.

[7] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes." *Proc. IEEE Int. Comm. Conf.*, (Geneva, Switzerland, 1993), pp. 1064–1070.

[8] A. R. Calderbank, G. D. Forney, Jr., and A. Vardy, "Minimal tail-biting trellises: the Golay code and more." submitted to *IEEE Trans. Inform. Theory.*

[9] G. D. Forney, Jr., F. R. Kschischang, and B. Marcus, "Iterative decoding of tail-biting trellises," presented at 1998 Information Theory Workshop. San Diego: Feb. 9–11, 1998, pp. 11-12.

[10] G. D. Forney, Jr., F. R. Kschischang, and A. Reznik, "The effective weight of pseudocodewords for codes defined on graphs with cycles on AWGN channels," Preprint.

[11] D. Lind and B. Marcus, *Symbolic Dynamics and Coding.* Cambridge, England:Cambridge University Press, 1995.

[12] R. Köetter, and A. Vardy, personal communication.

[13] G. Solomon and H. C. A. Van Tilborg, "A connection between block and convolutional codes." *SIAM J. APPL. MATH.*, pp. 358–367, October 1979.

[14] Y. Weiss, "Correctness of local probability propagation in graphical models with loops", submitted to *Neural Computation.*

[15] N. Wiberg, *Codes and Decoding on General Graphs.* Ph.D. dissertation, Department of Electrical Engineering, U. Linköping, Sweden (April 1996).

Turbo Codes with Short Frame Lengths*

Xia Lou
Department of Electrical Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0511
xlou@epicac.unl.edu

Lance C. Pérez
Department of Electrical Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0511
lperez@unl.edu

Turbo codes and other forms of low density parity check codes have been shown to achieve near capacity performance when long frame lengths are used. Unfortunately, long frame lengths result in decoding delays that are prohibitively long for many systems that offer voice service [1]. Here, the design and performance of turbo codes with short frame lengths, e.g. $N \leq 1000$, are considered by examining the weight distributions of the component codes. The effect of the weight distribution on performance is determined using the analytical techniques developed in [2]-[4] and by simulation.

Previous results [3, 4, 5] have shown that for large frame lengths, the performance of a turbo code is improved if the feedback polynomial of the constituent encoders is chosen to be primitive. This is due to the fact that primitive feedback polynomials generally maximize the minimum parity weight associated with input sequences of weight 2 and for large N these codewords dominate the performance. However, for small frame lengths input sequences of weight 2 are no longer the only input sequences of importance.

To illustrate this, consider the standard turbo code based on identical rate 1/2, memory $\nu = 4$, i.e. $(2,1,4)$, constituent encoders. The encoders are realized in systematic feedback form with parity check polynomials $h_1 = 35 = D^4 + D^3 + D^2 + 1$ and $h_0 = 23 = D^4 + D + 1$. In this case, the feedback polynomial $h_0 = 23$ is a primitive polynomial. It is well known that for large N the performance of this code is determined by input sequences of weight 2. In Figure 1, the contributions of different weight information sequences to the overall probability of error are shown for $N = 1000$ as a function of the bit error rate. (Note that here and in the remainder of this paper no puncturing is considered.) This figure clearly shows that inputs sequences of weight 3 contribute significantly to the error performance for all BERs and overwhelmingly dominate the performance for BERs less than 10^{-4}.

In designing an asymmetric turbo code for short frame lengths, it was endeavored to choose one encoder to provide a favorable distribution for input sequences of weight 2 and the other encoder to provide a favorable distribution for input sequences of weight 3. The former goal is met by using a primitive polynomial for the feedback polynomial. Thus, the first constituent encoder was chosen to be the $(2,1,4)$ code with polynomials $h_0 = 23$ and $h_1 = 35$. The second constituent encoder was chosen to be the $(2,1,2)$ code with parity polynomials $h_0 = 7$ and $h_1 = 5$. The contributions of the different weight information sequences to the overall probability of error for this asymmetric turbo code are shown in Figure 2 for $N = 1000$. This figure shows that for BERs greater than

---

*This research is supported in part by NSF Grant No. NCR-9714270 and the Center for Communications and Information Sciences at the University of Nebraska

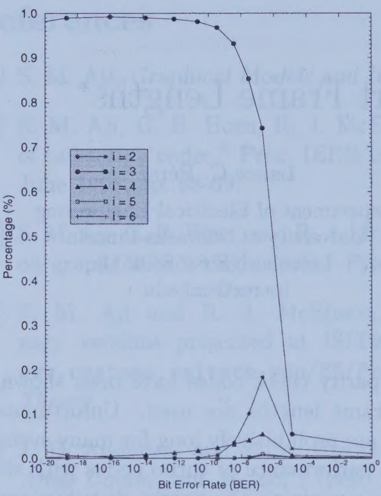


Figure 1: Performance as a function of input weight for a symmetric turbo code with memory $\nu = 4$ component codes.

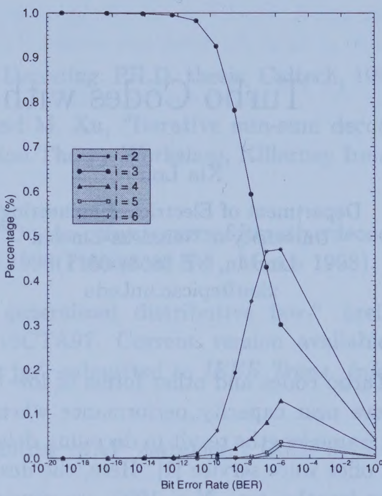


Figure 2: Performance as a function of input weight for an asymmetric turbo code with memory $\nu = 4$ and $\nu = 2$ component codes.

$\approx 10^{-6}$, the performance is now dominated by weight two input sequences. For BERs less than $\approx 10^{-6}$, the performance is once again dominated by weight 3 information sequences. The point at which this transition occurs is a function of the frame length N. Simulation results have shown that the asymmetric code outperforms the symmetric code in terms of both frame and bit error rate.

References

[1] P. Jung and M. NaBhan, "Performance Evaluation of Turbo Codes for Short Frame Transmission Systems", Electronics Letters, Vol. 30 N0. 2, pp. 111–113, Jan. 1994.

[2] S. Benedetto and G. Montorsi, "Design of Parallel Concatenated Convolutional Codes", *IEEE Trans. on Comm.*, Vol. 44, pp. 591–600, May 1996.

[3] D. Divsalar and F. Pollara, "On the Design of Turbo Codes", TDA Progress Report pp. 42–123, Nov. 1995.

[4] D. Divsalar, S. Dolinar, and F. Pollara, R. J. McEliece, "Transfer Function Bounds on the Performance of Turbo Codes", TDA Progress Report 42-122, pp. 44–55, Aug. 1995.

[5] L. C. Perez, J. Seghers, and D. J. Costello, Jr., "A Distance Spectrum Interpretation of Turbo Codes", *IEEE Trans. on Inform. Theory*, Vol. 42, No. 6, pp. 1698–1709, Oct. 1996.

An Algebraic Framework for Turbo Codes*

Brian M. Allen
University of Notre Dame, Mathematics Department
370 CCMB, Notre Dame, Indiana 46556-5683, USA
e-mail: ballen1@nd.edu

1 Introduction

Turbo codes were introduced by Berrou, Glavieux and Thitimajshima in 1993 [1]. Their idea of using parallel concatenation of recursive systematic convolutional codes (RSCs) with an interleaver was a major step in terms of achieving low bit error rates (BERs) at signal-to-noise ratios (SNRs) near the Shannon limit. A myriad of authors have since begun the task of analyzing why turbo codes are so effective. Work has been focused on the weight distribution, interleaver design, and the iterative soft-decision decoder. This paper will develop a mathematical framework that will enable further clarification and improvement in all these areas.

1.1 Turbo Codes

We restrict ourselves to only considering turbo codes where two parallel identical RSCs are used and no puncturing of the output bits is employed. With these assumptions, we give a brief overview of turbo encoding.

A turbo codeword consists of three parts: the information vector, the parity vector obtained from the RSC when the information vector is encoded, and a second parity vector obtained from the second RSC using a permuted version of the information vector.

1.2 The Input-State-Output Representation

These ideas were previously presented in [3]. They are included here for completeness. Let $\mathbb{F} = GF(q)$ be the Galois field with q elements and consider the matrices $A \in \mathbb{F}^{\delta\times\delta}$, $B \in \mathbb{F}^{\delta\times k}$, $C \in \mathbb{F}^{(n-k)\times\delta}$, and $D \in \mathbb{F}^{(n-k)\times k}$. A rate $\frac{k}{n}$ convolutional code $\mathcal{C}$ of complexity δ can be described by the linear system governed by the equations:

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t, \quad x_0 = 0 \\ y_t &= Cx_t + Du_t \end{aligned} \tag{1.1}$$

Hence the input to the encoder at time t is the *information vector*, u_t. The encoder creates the *parity vector*, y_t. The matrices (A, B, C, D) are called an input-state-output (ISO) representation of the convolutional code.

2 A New Description of Turbo Codes

In this section we will use the representation of a convolutional code presented in the previous section to obtain representations of turbo codes. In particular, we will obtain a generator matrix and parity check matrix description of turbo codes.

First, let us develop our turbo code. Denote the interleaver by S and its size by N. We choose our identical RSCs by specifying the matrices (A, B, C, D). For simplicity, we assume our RSCs are of rate 1/2. We will make the somewhat standard assumption that

---

*This research was supported in part by the University of Notre Dame Center for Applied Math and NSF grant DMS-96-10389.

after each N time intervals both RSCs will be back in the all zero state. This is possible given an appropriate choice of interleaver [2].

We can represent each N-block of output from the RSC as follows using the ISO representation. For $\tau \geq 0$ we have:

$$\begin{pmatrix} y_{\tau N} \\ y_{\tau N+1} \\ \vdots \\ \vdots \\ y_{\tau N+N-1} \end{pmatrix} = \begin{pmatrix} D & 0 & \cdots & & 0 \\ CB & D & \ddots & & \vdots \\ CAB & CB & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & 0 \\ CA^{N-2}B & CA^{N-3}B & \cdots & CB & D \end{pmatrix} \begin{pmatrix} u_{\tau N} \\ u_{\tau N+1} \\ \vdots \\ \vdots \\ u_{\tau N+N-1} \end{pmatrix}. \tag{2.1}$$

Denote the large matrix on the right hand side of the above equation by M. We are now ready to present a series of propositions which provide a parity check matrix, pseudo-generator and generator matrix description of the turbo code. The proofs of these propositions are straightforward from the ISO representation.

Proposition 2.1 *Any sequence in $\mathbb{F}$ is a codeword of the turbo code if and only if it is a sequence of valid N-blocks. A vector $C \in \mathbb{F}^{3N}$ is a valid N-block if and only if*

$$\begin{bmatrix} \Phi^N(A,B) & 0 & 0 \\ M & -I & 0 \\ MS & 0 & -I \end{bmatrix} C = 0$$

Proposition 2.2 *If $u \in \mathbb{F}^N$ and $u \in ker\, \Phi^N(A,B)$, then the image Gu, where $G = [I \;\; M' \;\; (MS)']'$, is a valid N-block of the turbo code.*

This pseudo-generator description is not satisfying since the input, u, is constrained. However, for reasonable A and B, Cayley's theorem implies that $\Phi^N(A,B)u = 0$ if and only if $P_A(x)|u(x)$, where $P_A(x)$ is the minimum polynomial of the matrix A, and $u(x)$ is the standard polynomial representation of the information vector u. We define the matrix P, of size $N \times N - \delta$, with the property that the image of P is precisely the kernel of $\Phi^N(A,B)$, by letting P be the matrix whose columns are the coefficients of $P_A(x)$ in the usual sliding generator matrix fashion.

Proposition 2.3 *The redefined matrix $G = [\, P' \;\; (MP)' \;\; (MSP)' \,]'$ is a generator matrix for valid N-blocks of the turbo code.*

3 Conclusions

The main issue of this paper was to present turbo codes in a mathematical framework that was not present in earlier works on the subject. Using some linear systems results, a framework for turbo codes, including parity check and generator matrix descriptions was obtained.

References

[1] C. Berrou, A. Glavieux, and P. Thitimajshima. Near Shannon limit error-correcting coding and decoding: Turbo-codes. In *Proc. of IEEE Int. Conference on Communication*, pages 1064–1070, Geneva, Switzerland, May 1993.

[2] W.J. Blackert, E.K. Hall, and S.G. Wilson. Turbo code termination and interleaver conditions. *Electronics Letters*, 31(24):2082–2083, 1995.

[3] E.V. York. *Algebraic Description and Construction of Error Correcting Codes, a Systems Theory Point of View.* PhD thesis, University of Notre Dame, 1997. Available at http://www.nd.edu/~rosen/preprints.html.

Turbo Codes for APD-Detected PPM*

Jon Hamkins and Meera Srinivasan

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109-8099
{hamkins,meera.srinivasan}@jpl.nasa.gov

Abstract

This paper introduces a method to use binary turbo codes with high order pulse position modulation (PPM) and an avalanche photodiode (APD) detector typical of an optical communications system. In general, the turbo codes outperform the Reed-Solomon codes usually recommended for this application. The simulation results are obtained from an accurate modeling of the APD output statistics, instead of Gaussian approximations which are often used but which also can over-predict or under-predict symbol error probabilities.

1 Introduction

Optical communication systems currently under development for the deep space channel use high order pulse position modulation (PPM) as the data modulation format. Information bits are encoded, modulated using M-ary PPM, transmitted optically, and received by an avalanche photodiode (APD) detector. For each PPM symbol, the APD produces M soft outputs corresponding to the output electron count for each slot duration.

The exact distribution of output electrons from the APD has been given in [4,8], but is cumbersome to use, and may be approximated very closely by the Webb density function [13]. In addition, the follow-on electronics contribute additive Gaussian thermal noise to the APD shot noise. A simpler analysis involves modeling both APD shot noise and Gaussian thermal noise as a single Gaussian process. In [5], it was shown that for 4-ary PPM, the purely Gaussian approximation overestimates PPM symbol error probability in the region $P_e < 0.01$. In this paper we provide symbol error probability calculations using both the accurate Webb plus Gaussian model as well as the strictly Gaussian approximation for 256-ary PPM, and show that the Gaussian approximation may either over-predict or under-predict symbol error probability by varying degrees depending upon the levels of background and signal light levels. An upper bound using the Webb plus Gaussian mixture density is provided as a more uniform method of estimating PPM symbol error probability.

Using the Webb model of APD shot noise, we can accurately simulate the performance of a coded system. We introduce a binary turbo coding scheme for high order PPM, and

---

*The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

compare it to equivalent rate Reed-Solomon codes. Previous work on coded PPM signaling (see, e.g., [9]) has concentrated on Reed-Solomon codes because their alphabet sizes are easily matched to the PPM order, resulting in a one-to-one correspondence between code symbols and PPM symbols. Since every Reed-Solomon code is a maximum distance code, it might seem that no more powerful code could be found for this application.

However, a disadvantage of a Reed-Solomon code is that the decoder does not use the soft detector outputs directly, but instead operates on decoded PPM symbols and erasures. This is an inherent limitation of Reed-Solomon codes, because there is no method to modify Reed-Solomon codes to make full use of the soft information. We thus have the motivation for using turbo codes, which are already known to have outstanding performance in other applications and can operate on the soft counts directly rather than having to use explicit PPM symbol decisions.

This paper makes three new contributions: (1) Uncoded PPM symbol error probability calculations for 256-ary PPM, (2) A binary turbo coding scheme for high order PPM that fully utilizes soft information, and (3) A simulation of end-to-end BER performance for baseline Reed-Solomon codes and binary turbo codes using an accurate statistical model of the APD detector in the absence of thermal noise.

2 APD Receiver Modeling

The average number of photons absorbed by an APD illuminated with total optical power $P(t)$ in T_s seconds can be expressed as

$$\bar{n} = \frac{\eta}{h\nu} \int_0^{T_s} P(t)dt \tag{1}$$

where h is Planck's constant, ν is the optical frequency, and η is the detector's quantum efficiency, defined as the ratio of absorbed to incident photons. The actual number of photons absorbed, n, is a Poisson distributed random variable with rate $\bar{n}$. The conditional probability that an APD generates m electrons in response to exactly n absorbed photons, $m \geq n$, $n > 0$, has been shown by McIntyre to be [8]

$$p(m|n) = \frac{n\Gamma\left(\frac{m}{1-k}+1\right)}{m(m-n)!\Gamma\left(\frac{km}{1-k}+n+1\right)} \left[\frac{1+k(G-1)}{G}\right]^{n+km/(1-k)} \left[\frac{(1-k)(G-1)}{G}\right]^{m-n} \tag{2}$$

where G is the average APD gain, and k is the ionization ratio. Averaging over the exact number of absorbed photons, we obtain

$$p(m|\bar{n}) = \sum_{n=1}^{m} p(m|n)\frac{\bar{n}^n}{n!}e^{-\bar{n}}, \qquad m \geq 1 \tag{3}$$

as the distribution of APD output electrons generated over T_s seconds given the mean number of absorbed photons in that interval. If no photons are absorbed, then no electrons can be generated according to this model. Conversely, no electrons are generated only if no photons have been absorbed. Hence, $p(m=0|\bar{n}) = p(n=0|\bar{n}) = e^{-\bar{n}}$.

An approximation to Equation (3) has been derived by Webb [13] and provides a much simpler expression for the density of m in response to the mean number of photons absorbed:

$$p_w(m|\bar{n}) = \frac{1}{\sqrt{2\pi\bar{n}G^2F}\left(1+\frac{m-G\bar{n}}{\bar{n}GF/(F-1)}\right)^{3/2}} \exp\left(-\frac{(m-G\bar{n})^2}{2\bar{n}G^2F\left(1+\frac{m-G\bar{n}}{\bar{n}GF/(F-1)}\right)}\right) \tag{4}$$

where $F = kG + (2 - 1/G)(1 - k)$. This density is a continuous function that is defined for $m > -\lambda G/(F-1)$, and integrates to one in that range. The discrete probabilities in (3) are approximated by evaluating (4) at integer values of m. Note that (4) is defined for negative values of m, provided $m > -\lambda G/(F-1)$, even though negative m has no physical meaning. In the numerical evaluations presented here, we calculate the sum of the Webb density values for $m \geq 1$, and assign to $p(m = 0|\bar{n})$ the value of this sum subtracted from one.

Added to the random number of APD output electrons is an independent Gaussian thermal noise charge from the follow-on electronics. The total charge is then integrated over each slot time T_s, resulting in a vector of M independent observables for each received PPM word. It was proven in [12] that given these observables, the maximum likelihood detector structure consists of choosing the PPM symbol corresponding to the slot with the maximum accumulated charge value. If $\bar{n}_b$ and $\bar{n}_s$ are the mean number of absorbed background photons per slot and the mean number of absorbed signal photons per pulse, respectively, the M-ary PPM symbol error probability is

$$P_e(M) = 1 - \int_{-\infty}^{\infty} p(x|\bar{n}_b + \bar{n}_s) \left[\int_{-\infty}^{x} p(y|\bar{n}_b) dy \right]^{M-1} dx, \tag{5}$$

where $p(x|\bar{n})$ is the probability density function for the slot statistic given $\bar{n}$ mean number of absorbed photons over the slot duration. Since the slot statistic is a random variable consisting of the sum of independent Webb and Gaussian random variables, its density function is the convolution of the individual densities, and may be written as

$$p(x|\bar{n}) = \sum_{m=0}^{\infty} \phi(x, \mu_m, \sigma^2) p_w(m|\bar{n}), \tag{6}$$

where $\phi(x, \mu_m, \sigma^2)$ is the Gaussian density function with mean $\mu_m = mq + I_s T_s$ and variance $\sigma^2 = \left(2qI_s + \frac{4\kappa T}{R}\right) BT_s^2$, as given in [5]. Here, q is the electron charge, κ is Boltzmann's constant, T is the equivalent noise temperature, B is the single-sided noise bandwidth, and I_s is the APD surface leakage current. Note that the APD surface leakage current is not multiplied by the APD gain and is modeled here as a constant DC current. The APD dark current, on the other hand, is multiplied by the APD gain and is modeled as part of the background radiation, i.e., it is incorporated into the value of $\bar{n}_b$.

An approach that is sometimes used to simplify calculation of PPM symbol error probabilities is to model the density of the APD output electron charge as Gaussian with mean $qG\bar{n}$ and variance $q^2G^2F\bar{n}$. Then the slot statistic consisting of the sum of APD output electrons and amplifier thermal noise is also Gaussian, and has mean $\mu = qG\bar{n} + I_sT_s$ and variance $\sigma^2 = \left[2q^2G^2F\bar{n} + qI_sT_s + \frac{4\kappa TT_s}{R}\right] BT_s$. Although simple, this approximation does not yield accurate results over all regions of interest, as previously shown in [5] for 4-PPM and as we demonstrate here for 256-PPM.

3 Uncoded PPM Symbol Error Probabilities

Substituting (6) into (5), the PPM symbol error probability may be expressed as

$$\begin{aligned} P_e(M) &= 1 - \int_{-\infty}^{\infty} \sum_{m=0}^{\infty} \phi(x, \mu_m, \sigma^2) p_w(m|\bar{n}_b + \bar{n}_s) \left[\int_{-\infty}^{x} \sum_{n=0}^{\infty} \phi(y, \mu_n, \sigma^2) p_w(n|\bar{n}_b) dy \right]^{M-1} dx \\ &= \sum_{m=0}^{\infty} p_w(m|\bar{n}_b + \bar{n}_s) \int_{-\infty}^{\infty} \phi(x, \mu_m, \sigma^2) \left(1 - \left[\sum_{n=0}^{\infty} p_w(n|\bar{n}_b) \Phi\left(\frac{x - \mu_n}{\sigma}\right) \right]^{M-1} \right) dx, \end{aligned} \tag{7}$$

where $\Phi(x)$ is the Gaussian distribution function $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$. The expression in (7) is difficult and time-consuming to evaluate accurately, so an upper bound is calculated instead. Although the union bound is commonly used for such M-ary orthogonal signaling problems, a tighter bound was derived by Hughes [7] and is applied here. By observing that the function $f(u) = 1 - [1-u]^{M-1}$ is concave for $u \in [0,1]$, Jensens's inequality gives us $E[f(u)] \leq f(E[u])$. Letting $u = \int_x^{\infty} p(y|\bar{n}_b)dy$ and taking the expectation over x with density function $p(x|\bar{n}_b + \bar{n}_s)$, we can then show that

$$\begin{aligned} P_e(M) &= \int_{-\infty}^{\infty} p(x|\bar{n}_b + \bar{n}_s) \left(1 - \left[1 - \int_x^{\infty} p(y|\bar{n}_b)dy \right]^{M-1} \right) dx \\ &\leq 1 - \left[1 - \int_{-\infty}^{\infty} p(x|\bar{n}_b + \bar{n}_s) \int_x^{\infty} p(y|\bar{n}_b)dydx \right]^{M-1}. \end{aligned} \qquad (8)$$

The double integral within the brackets on the right hand side of (8) is simply the probability that a given background slot has a greater electron count than the signal plus background slot. This may also be interpreted as the symbol error probability for equiprobable binary PPM signaling, denoted as $P_e(2)$, and given by

$$P_e(2) = \sum_{m=0}^{\infty} p_w(m|\bar{n}_b + \bar{n}_s) \int_{-\infty}^{\infty} \phi(x, \mu_m, \sigma^2) \sum_{n=0}^{\infty} p_w(n|\bar{n}_b) \Phi\left(-\frac{x-\mu_n}{\sigma}\right) dx, \qquad (9)$$

so that the upper bound on the M-ary PPM symbol error probability may now be expressed as $P_e(M) \leq 1 - (1 - P_e(2))^{M-1}$.

Although the upper bound in (8) may be calculated relatively easily, it does require repeated evaluation of the Webb density and nested summation and integration. Use of the Gaussian approximation to the APD statistics saves computational time. Using this Gaussian approximation, the M-ary PPM symbol error probability is given by

$$P_{e,\text{Gaussian}}(M) = \int_{-\infty}^{\infty} \phi(x, \mu_s, \sigma_s^2) \left[1 - \Phi\left(\frac{x-\mu_b}{\sigma_b}\right) \right]^{M-1} dx, \qquad (10)$$

where $\mu_b = qG\bar{n}_b + I_sT_s$, $\mu_s = qG(\bar{n}_b + \bar{n}_s) + I_sT_s$, $\sigma_b^2 = \left[2q^2G^2F\bar{n}_b + qI_sT_s + \frac{4\kappa TT_s}{R}\right] BT_s$, and $\sigma_s^2 = \left[2q^2G^2F(\bar{n}_b + \bar{n}_s) + qI_sT_s + \frac{4\kappa TT_s}{R}\right] BT_s$ [5].

The two expressions for evaluating PPM symbol error probability given in equations (8) and (10) were calculated and compared using the following parameters: $M = 256$, $T_s = 2 \times 10^{-8}$ seconds, $k = 0.007$, $I_s = 2 \times 10^{-9}$ amps, $T = 300°$ K, $R = 146650\Omega$, and $B = 1/2T_s = 2.5 \times 10^7$ Hz ($1/2T_s$ is the noise equivalent bandwidth for an ideal integrator over duration T_s). Calculations were performed for several different values of $\bar{n}_b$ and $\bar{n}_s$. The APD gain parameter G was optimized for each pair $\{\bar{n}_b, \bar{n}_s\}$ by taking on the value resulting in minimum PPM symbol error probability. The optimization was performed separately for the Webb and the Gaussian models, resulting in different optimum gains for each model. In addition to numerical evaluation, Monte Carlo simulations of the receiver performance under the Webb model were performed in order to obtain more accurate results for lower SNR regions than those provided by the upper bound (8). In the Monte Carlo simulations, the Webb deviates needed to simulate APD output were generated using the inclusion-exclusion principle [10] following the method outlined in [11].

Fig.s 1-2 show plots of P_e versus $\bar{n}_s$ for three particular values $\bar{n}_b$: 0.01, 1.0, and 100.0. The three curves shown are the error probability for the Webb model obtained through Monte Carlo simulation, the upper bound for the Webb model as given by (8), and the error probability for the Gaussian approximation as given by (10). From these figures we see that the upper bound (8) approaches the exact error probability as given by the Webb simulation as P_e becomes small. In Fig. 1, the Gaussian approximation consistently overbounds the simulation error probability, whereas in Fig. 2, the Gaussian curve first underbounds the Webb curve for lower $\bar{n}_s$ and then crosses over and overbounds the Webb curve as $\bar{n}_s$ increases. The cross-over points occur at increasing values of signal photon level as the background photon level increases.

4 Turbo-Coded M-ary PPM

4.1 Turbo Encoder

A rate 1/3 turbo encoder is shown in Fig. 3. It consists of two rate 1/2 constituent codes and an interleaver of length N. The input $\mathbf{u} = (u_1, ..., u_N)$ is a block of independent, uniformly distributed information bits. We use the trellis termination method in [6], which takes m additional steps to return the encoders to the all zeroes state.

In this paper, the constituent codes shall be rate 1/2, recursive, systematic convolutional codes. The three outputs of the rate 1/3 turbo encoder include the systematic bit x_0, and the two parity bits x_{1p} and x_{2p}. The permuted systematic bit from the second convolutional code is not transmitted across the channel. Thus, a total of $n = 3(N+m)$ bits are transmitted for the rate 1/3 code. The bits are grouped into M-ary PPM symbols. For concreteness, in this paper $M = 256$.

4.2 Turbo Decoder

We use a turbo decoder structure and notation as in [2], shown in Fig. 4. Although this structure was designed as a turbo decoder for a binary AWGN channel, part of its beauty is that each soft-input soft-output (SISO) module need not know anything about the channel in order to operate— its computation is determined completely by the a priori distribution $P(\mathbf{c}; I)$ and knowledge of the code structure. Thus, given the a priori probability distributions, the turbo decoder operates in the usual way. Details are contained in [1–3, 6].

To properly initialize the turbo decoder, we must determine the a priori probability of codewords for each constituent code.[1] Let $\mathbf{c} = (c_0, c_1)$ denote a possible constituent code codeword. As in [2], we define $P(\mathbf{c}; I) = \{P_k(\mathbf{c}; I)\}$, for $k = 1, \ldots, N+m$, where for SISO 1

$$P_k(\mathbf{c}; I) = \Pr[\text{codeword } (x_0, x_{1p}) \text{ is } \mathbf{c} \text{ at time } k| \text{ all soft counts from all symbols}], \tag{11}$$

and for SISO 2

$$P_k(\mathbf{c}; I) = \Pr[\text{codeword } (\pi(x_0), x_{2p}) \text{ is } \mathbf{c} \text{ at time } k| \text{ all soft counts from all symbols}]. \tag{12}$$

---

[1] By "a priori" probability, we mean that there is conditioning on the the receiver decision statistics, but no conditioning on the turbo code structure, i.e., for the purposes of computing $P(\mathbf{c}; I)$, the components of $\mathbf{c}$ are assumed to be independent and uniformly distributed. This the usual meaning in the the context of turbo codes; in other contexts, conditioning on receiver statistics would be termed a posteriori.

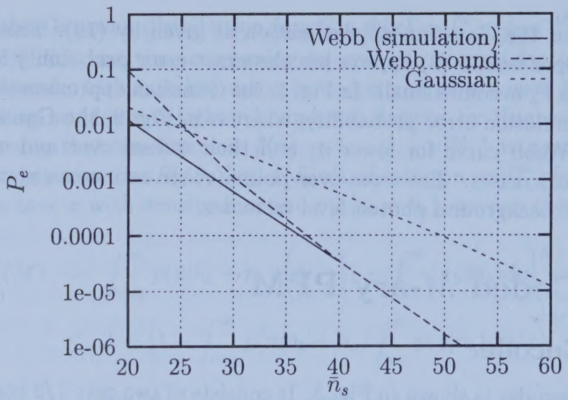


Figure 1: Comparison of error probabilities using Webb and Gaussian models, $\bar{n}_b = 0.01$, optimized gain.

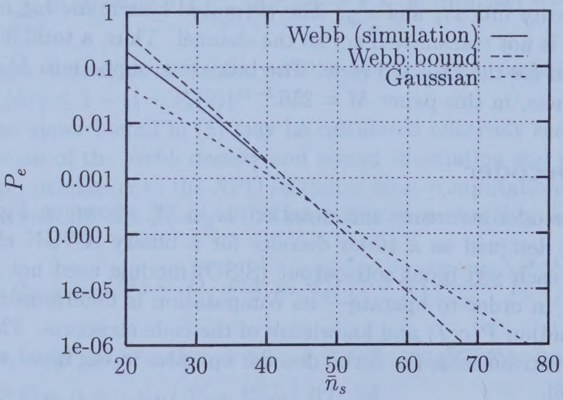


Figure 2: Comparison of error probabilities using Webb and Gaussian models, $\bar{n}_b = 1$, optimized gain.

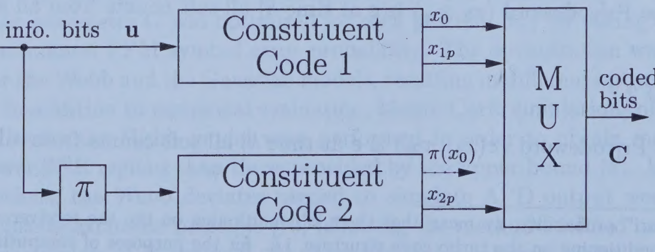

info. bits **u**
Constituent Code 1
x_0
x_{1p}
π
Constituent Code 2
$\pi(x_0)$
x_{2p}
M U X
coded bits
c

Figure 3: Rate 1/3 Turbo Encoder.

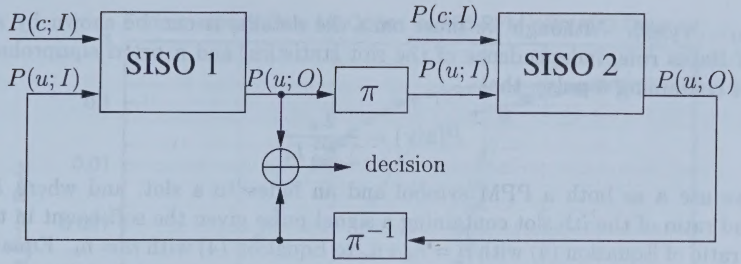


Figure 4: Turbo decoder.

Note that all soft counts in the conditioning in Equations (11) and (12) are irrelevant to computing $P_k(\mathbf{c};I)$ except those from PPM symbols that contain the kth constituent codeword bits. That is, a soft count does not give information about (x_0, x_{1p}) unless it is a soft count from a PPM symbol that depends on x_0 or x_{1p}.

To compute $P_k(\mathbf{c};I)$, there are two cases to consider. We consider SISO 1 here; the analysis is identical for SISO 2, except that (x_o, x_{1p}) is replaced by $(\pi(x_0), x_{2p})$.

Case 1: *The bits (x_0, x_{1p}) which form the kth constituent codeword are grouped into two separate PPM symbol epochs s and t.*

In this case, let x_0 be the lth bit of the sth PPM symbol, let x_{1p} be the mth bit of the tth PPM symbol, and let $\mathbf{y} = (y_0, \ldots, y_{255})$ and $\mathbf{z} = (z_0, \ldots, z_{255})$ denote the two sets of soft counts from the two transmitted PPM symbols. Then

$$\begin{align}
P_k(\mathbf{c};I) &= P_k((c_0,c_1);I) \notag\\
&= P(\text{bit } l \text{ of } s\text{th PPM symbol is } c_0|\mathbf{y}) \tag{13}\\
&\qquad \cdot P(\text{bit } m \text{ of } t\text{th PPM symbol is } c_1|\mathbf{z}) \tag{14}\\
&= \left(\sum_{\substack{\mathbf{a}=(a_1,\ldots,a_8)\\ a_l=c_0}} P(\mathbf{a}|\mathbf{y})\right)\cdot\left(\sum_{\substack{\mathbf{a}=(a_1,\ldots,a_8)\\ a_m=c_1}} P(\mathbf{a}|\mathbf{z})\right), \tag{15}
\end{align}$$

where Equation (14) follows by the independence of the soft counts from one PPM symbol to the next, and where in Equation (15) $\mathbf{a}$ is the 8-bit vector defining a 256-PPM symbol.

Case 2: *The bits (x_0, x_{1p}) which form the kth constituent codeword are grouped into the same PPM symbol epoch s.*

In this case, let x_0 be the lth bit of the sth PPM symbol, let x_{1p} be the mth bit of the sth PPM symbol, and let $\mathbf{y} = (y_0, \ldots, y_{255})$ denote the set of soft counts from the sth PPM symbol. Then

$$\begin{align}
P_k(\mathbf{c};I) &= P_k((c_0,c_1);I) \notag\\
&= P(\text{bit } l \text{ and } m \text{ of } s\text{th PPM symbol is } c_0 \text{ and } c_1\text{, resp.}|\mathbf{y}) \notag\\
&= \sum_{\substack{\mathbf{a}=(a_1,\ldots,a_8)\\ a_l=c_0\\ a_m=c_1}} P(\mathbf{a}|\mathbf{y}). \tag{16}
\end{align}$$

In either Case 1 or Case 2, we need to compute the probability that a given 256-PPM symbol $\mathbf{a} = (a_1, \ldots, a_8)$ was transmitted, given its corresponding set of 256 soft counts

$\mathbf{y} = (y_0, \ldots, y_{255})$. Although we must omit the details, it can be shown by standard usage of Bayes rule, independence of the slot statistics, and a priori equiprobability of each slot containing a pulse, that

$$P(\mathbf{a}|\mathbf{y}) = \frac{L_{\mathbf{a}}}{\sum_{i=0}^{255} L_i}, \tag{17}$$

where we use $\mathbf{a}$ as both a PPM symbol and an index to a slot, and where L_i is the likelihood ratio of the ith slot containing a signal pulse given the soft count in that slot, i.e., the ratio of Equation (4) with $\bar{n} = \bar{n}_b + \bar{n}_s$ to Equation (4) with $\bar{n} = \bar{n}_b$. Equation (17) may be used in Equations (15) and (16) to compute the a priori probabilities needed for the SISO modules.

5 Performance Simulations

Using the method of the previous section, simulations were conducted for rate 1/3 turbo coded 256-PPM and a typical APD detector. The constituent codes had either 4 states and identical generators of $(5/7)_{oct}$, or 8 states and identical generators of $(17/15)_{oct}$. The interleaver had length 16382 and represented a random permutation. The signal varied from $\bar{n}_s = 20$ to 30 photons/slot, and the background intensity was fixed at $\bar{n}_b = 10$ photons/slot. We used the same APD parameters as in Section 3, except that the gain was fixed at 80, and we used the shot noise limited case, i.e., we assumed negligible additive thermal noise in the APD. The turbo decoder performed 10 iterations before bit decisions were made. The simulations were continued until 100 error-containing blocks were processed. This typically involved the simulation of 10,000 or more bit errors.

Reed-Solomon performance was determined by simulating the soft counts using a program to generate Webb deviates, computing the resulting PPM symbol error rate (SER) when the maximum likelihood rule is used [12], and using a formula for the performance of Reed-Solomon code based on that SER. Erasures were not considered.

Fig. 5 shows the performance of rate 1/3 coded 256-PPM as a function of input SER. (Fig. 6 shows the same simulation data, as a function of $\bar{n}_s$.) Since the turbo decoder does not make an explicit determination of PPM symbols, the input SER is technically not defined for turbo codes; the SER used in the plot is that which would have occurred had the maximum likelihood rule been used to make PPM symbol decisions. The three curves for turbo codes represent the 4-state noninterleaved PPM, 4 state interleaved PPM, and 8 state interleaved PPM simulations. By noninterleaved PPM, we mean that the bits at the output of the turbo encoder are grouped into 8-bit blocks directly and PPM modulated. This order is $(x_0, x_{1p}, x_{2p}, x_0, x_{1p}, x_{2p}, \ldots, x_0, x_{1p}, x_{2p})$. For the interleaved PPM, the order of the bits has been partially randomized before grouping into PPM symbols. The effect of the interleaver before the modulator is to spread out related bits into different PPM symbols.

Acknowledgements: The authors acknowledge Dariush Divsalar and Victor Vilnrotter for helpful discussions about turbo codes and the APD detector.

References

[1] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Trans. Inform. Theory*, IT-20:284–287, March 1974.

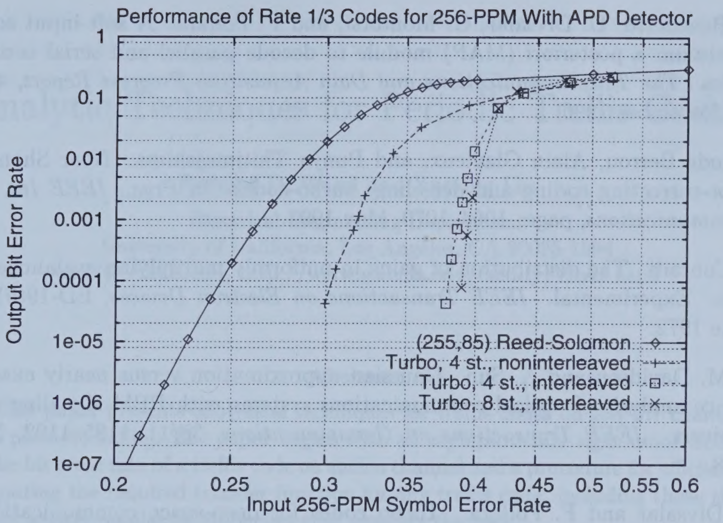


Figure 5: Performance of rate 1/3 codes as a function of input SER.

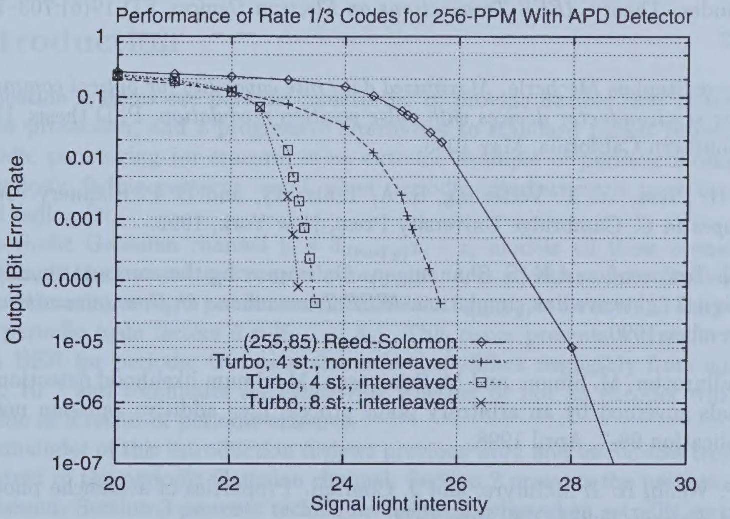


Figure 6: Performance of rate 1/3 codes as a function of signal intensity.

[2] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara. A soft-input soft-output maximum a posteriori (MAP) module to decode parallel and serial concatenated codes. *The Telecommunications and Data Acquisition Progress Report*, 42(127):1–20, November 1996.

[3] Claude Berrou, Alain Glavieux, and Punya Thitimajshima. Near Shannon limit error-correcting coding and decoding: turbo-codes. In *Proc., IEEE Int. Conf. on Communications*, pages 1064–1070, May 1993.

[4] J. Conradi. The distribution of gains in uniformly multiplying avalanche photodiodes: Experimental. *IEEE Transactions on Electron Devices*, ED-19(6):713–718, June 1972.

[5] F. M. Davidson and X. Sun. Gaussian approximation versus nearly exact performance analysis of optical communications systems with PPM signaling and APD receivers. *IEEE Transactions on Communications*, 36(11):1185–1192, November 1988.

[6] D. Divsalar and F. Pollara. Turbo codes for deep-space communications. *The Telecommunications and Data Acquisition Progress Report*, 42(120):29–39, February 1995.

[7] L. Hughes. A simple upper bound on the error probability for orthogonal signals in white noise. *IEEE Transactions on Communications*, COM-40(4):670, April 1992.

[8] R. J. McIntyre. The distribution of gains in uniformly multiplying avalanche photodiodes: Theory. *IEEE Transactions on Electron Devices*, ED-19(6):703–713, June 1972.

[9] George Stephen Mecherle. *Maximized data rate capability for optical communication using semiconductor devices with pulse position modulation.* PhD thesis, University of Southern California, May 1986.

[10] W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery. *Numerical Recipes in C.* Cambridge University Press, New York, 1992.

[11] J. K. Townsend and K. S. Shanmugan. On improving the computational efficiency of digital lightwave link simulation. *IEEE Transactions on Communications*, 38(11), November 1990.

[12] V. Vilnrotter, M. Simon, and M. Srinivasan. Maximum likelihood detection of PPM signals governed by an arbitrary point process plus additive gaussian noise. JPL Publication 98-7, April 1998.

[13] P. P. Webb, R. J. McIntyre, and J. Conradi. Properties of avalanche photodiodes. *RCA Review*, 35:234–278, June 1974.

Analytic Techniques for Periodic Trellis Codes*

Richard D. Wesel and Xueting Liu

Electrical Engineering Department
University of California, Los Angeles, CA 90095-1594
wesel@ee.ucla.edu, http://www.ee.ucla.edu/faculty/Wesel.html

Abstract

This paper presents analytical techniques for trellis codes on AWGN channels with periodic fading. The paper begins with an asymptotically tight union bound on the bit error rate of a trellis code on such a channel and a procedure for efficiently computing the required transfer function for any trellis code, including those that do not satisfy regularity properties that simplify the task.

Periodic erasures can cause an encoder that is non-catastrophic for the regular AWGN channel to display catastrophic behavior. Two techniques identify erasure patterns that will cause this behavior. One technique applies the traditional catastrophic test to a strictly equivalent multiple trellis code after removing columns corresponding to erased symbols. The second technique efficiently searches for zero distance loops in the state diagram.

1 Introduction

Communication systems use periodic puncturing to provide flexible rate variation, unequal error protection, and a progressive alternative to standard packet retransmission [1]. Periodic puncturing (or erasure) is an extreme example of periodic fading. More general periodic fading patterns result when periodic interleaving is used to mitigate correlated fading [2].

The periodic Gaussian channel $y_i = \tilde{a}_{i(\text{mod } p)} x_i + z_i$ models all these scenarios. In this model x_i is the transmitted constellation point, z_i is additive white Gaussian noise (AWGN) with variance $N_0/2$ per dimension, and each $\tilde{a}_{i(\text{mod } p)}$ is an entry in the p-element vector of periodic scale factors $\tilde{a} = [\tilde{a}_1 \ \ldots \ \tilde{a}_p]$. This paper presents a transfer function bound on BER for periodic Gaussian channels that differs negligibly from simulation for BER$< 10^{-4}$ and techniques for identifying whether or not an encoder will become catastrophic as a result of periodic erasures.

The remainder of this introduction reviews previous work and introduces trellis codes in the context of the periodic Gaussian channel. Section 2 presents the periodic transfer function bound. Section 3 presents techniques for identifying when a trellis encoder will have catastrophic error events as a result of periodic erasures. Section 4 concludes the paper.

The first work on trellis code design and analysis for periodic channels was the design of 2-state and 4-state trellis codes for $p = 2$ by Leung et al. [3] for coding across distinct

---

*This research is supported in part by NSF CAREER Award #9733089 and grants from the Xetron Corporation, Rockwell Semiconductor Systems, and Pacific Bell.

narrow-band Rayleigh fading channels. Also in [3], Leung et al. presented both analytical and numerical transfer function bounds for BER on Rayleigh fading channels. In another important paper, Lapidoth identified a computational technique in [4] that efficiently computes the bit error rate (BER) of convolutional codes over noiseless channels with periodic erasures.

More recently, Wesel et al. [2, 5, 6] and Knopp et al. [7] have designed trellis codes with more states and larger periods than those in [3] using periodic effective code length and periodic product distance metrics. Wesel et al. also designed trellis codes specifically for periodic erasures according to a residual Euclidean distance metric [8].

Any discussion of trellis codes on periodic Gaussian channels begins with the recognition that sub-sampled groups of symbols see identical fading. Let the normalized symbol-wise squared Euclidean distance be $d_i^2 = (\hat{x}_i - x_i)^2/\mathcal{E}_x$ where x_i and $\hat{x}_i$ are the correct and incorrect constellation points associated with the ith symbol of a trellis error event $x \to \hat{x}$. The periodic Gaussian channel scales the distances with the same index modulo p by the same scale factor $\tilde{a}_i$. Define the periodic squared distance $\tilde{d}_i^2$ for $i \in \{1, 2, \ldots p\}$ as the sum of the distances scaled by the same $\tilde{a}_i$:

$$\tilde{d}_i^2 = \sum_{m=0}^{\infty} d_{i+mp}^2 \,. \tag{1}$$

The p values of $\tilde{d}_i^2$ comprise the periodic squared distance vector $\tilde{d}^2$:

$$\tilde{d}^2 = \left[\tilde{d}_1^2 \;\; \tilde{d}_2^2 \;\; \ldots \;\; \tilde{d}_p^2\right] . \tag{2}$$

Define $\tilde{a}^2$ to be the p-element vector $\left[|\tilde{a}_1|^2 \;\; \ldots \;\; |\tilde{a}_p|^2\right]$. For a specified $\tilde{a}$, the pairwise error probability for the trellis error event $x \to \hat{x}$ is

$$P(x \to \hat{x}) = Q\left(\sqrt{\frac{\mathcal{E}_x\langle \tilde{a}^2, \tilde{d}^2\rangle}{2N_0}}\right) , \tag{3}$$

where $Q(t) = \int_t^{\infty} e^{-u^2/2}du/\sqrt{2\pi}$, the expression $\langle \tilde{a}^2, \tilde{d}^2\rangle$ represents an inner product, and $\tilde{d}^2$ is the periodic distance vector for $x \to \hat{x}$. The residual Euclidean distance $\mathrm{RED}^2(\tilde{a})$,

$$\mathrm{RED}^2(\tilde{a}) = \min_{\tilde{d}^2} \langle \tilde{a}^2, \tilde{d}^2\rangle \,, \tag{4}$$

identifies the smallest $\langle \tilde{a}^2, \tilde{d}^2\rangle$, the one that maximizes (3).

2 Transfer Function Bit Error Rate Bound

For any vector $\tilde{a}$ of periodic scale factors and AWGN noise variance $N_0/2$, a periodic transfer function bound produces asymptotically tight bounds on bit error rate. Leung, Wilson, and Ketchum identified a generalized transfer function in [3] which they used for examining 2-state and 4-state periodic trellis codes with $p = 2$ on Rayleigh fading channels. A similar periodic transfer function is

$$T_p(W_1, \ldots, W_p, I) = \sum_{w_1 \ldots w_p} \sum_i N(w_1, \ldots, w_p, i) I^i \prod_{j=1}^{p} W_j^{w_j} \,, \tag{5}$$

where w_i is that component of the Euclidean distance of an error event that is scaled by $\tilde{a}_i$, i is the number of information bit errors associated with the error event, and $N(w_1, w_2, \ldots, w_p, i)$ is the expected number of error events beginning at any of the p symbols in a fixed period that cause i information bit errors and have Euclidean distances $w_1 \ldots w_p$ scaled by $a_i \ldots a_p$ respectively. The expectation is over the uniform distribution on valid transmitted constellation point sequences. With $p = 1$ (5) reduces to the standard transfer function $T(W, I)$ with coefficients $N(w, i)$.

Using $Q(\sqrt{x+y}) \leq e^{-y/2} Q(\sqrt{x})$, an asymptotically tight union bound on BER for an AWGN channel with periodic fading scale factors $a_i \ldots a_p$ is computed as

$$P_e \leq \frac{1}{kp} Q\left(\sqrt{\mathrm{RED}^2(\tilde{a})\mathcal{E}_x/2N_0}\right) e^{\mathrm{RED}^2 \mathcal{E}_x(\tilde{a})/4N_0} \left.\frac{\partial T_p(W_1, \ldots, W_p, I)}{\partial I}\right|_{I=1, W_j = e^{-|\tilde{a}_j|^2/4N_0}} . \quad (6)$$

The bound in (6) is essentially the same as one used in [3] for numerical integration over the p-dimensional Rayleigh density function. The bound is useful by itself to analyze performance for a particular $\tilde{a}$.

Three steps comprise the procedure to compute the bound in (6). The first step produces a minimal state diagram that accurately represents the input bit error and Euclidean distance behavior of the trellis code. The second step uses this state diagram to produce a periodic state diagram where each transition represents a period on the periodic Gaussian channel. The third step uses this periodic transfer function to produce the desired bound on BER.

2.1 Minimal state diagrams for non-regular trellis codes

Existing papers address the problem of obtaining a state diagram describing the input bit error and Euclidean distance behavior of a specified trellis code. For a trellis code with ν_e memory elements, valid state diagrams can always be constructed with $2^{2\nu_e}$ states using Biglieri's product-state technique [9]. If the trellis code satisfies the regularity conditions due to Zehavi & Wolf [10] or similar conditions due to Rouanne & Costello [11], the state diagram can be constructed with only 2^{ν_e} states. Benedetto, Mondin, and Montorsi unify these two regularity techniques in [12] and observe that almost all good trellis codes for AWGN channels satisfy one of these conditions allowing a 2^{ν_e}-state diagram.

Unfortunately, many good trellis codes for periodic Gaussian channels don't satisfy these regularity conditions. A technique that identifies a minimal state diagram for such non-regular trellis codes is absent from the existing literature. What follows is a general procedure to find this minimal trellis code state diagram. This procedure finds 2^{ν_e}-state diagrams for regular trellis codes as a special case.

We begin with a variation on Biglieri's product-state approach [9]. Instead of defining the product state as the pair (correct state, incorrect state) as in [9], we use the isomorphic pair (state error, encoder state). Let x, e_i, and e_o denote the binary constellation point label, input binary error, and output binary error respectively. Let σ_e denote the error state, i.e. the bitwise exclusive-or of the correct state and the incorrect state. Let σ_x denote the correct encoder state. The $2^{2\nu_e}$-state diagram has product states (σ_e, σ_x) and each transitions $(s_e, s_x) \rightarrow (t_e, t_x)$ labeled with

$$P(s_x \rightarrow t_x) \sum_{x} \sum_{(e_i, e_o | s_e \rightarrow t_e)} P(x | s_x \rightarrow t_x) W^{d_E^2(x, x \oplus e_o)} I^{d_H(e_i, 0)} . \quad (7)$$

The second summation is over input and output binary errors possible for the transition of $s_e \to t_e$, $d_E^2(x, x \oplus e_o)$ is the squared Euclidean distance between the constellation points labeled with x and $x \oplus e_o$, and d_H is Hamming distance. When there are no parallel paths in the trellis, there is only one nonzero term in this summation and $P(x|s_x \to t_x) = 1$. The overall product-state diagram is essentially the direct product of the σ_e state diagram and the σ_x state diagram with only the exponent of W involving both x and e.

A smaller state diagram will often suffice. Assign each constellation point to one of a set of equivalence classes $\{\mathcal{C}\}$ indexed by q where all $x \in \mathcal{C}_q$ share the same mapping of e_o to $d_E^2(x, x \oplus e_o)$. Define $d_E^2(q, e_o)$ to be equal to the value of $d_E^2(x, x \oplus e_o)$ for $x \in \mathcal{C}_q$. The transition label in (7) is equal to

$$P(s_x \to t_x) \sum_q \sum_{(e_i, e_o | s_e \to t_e)} P(q|s_x \to t_x) W^{d_E^2(q,e_o)} I^{d_H(e_i,0)}, \tag{8}$$

The σ_x state diagram has one role in the computation of $T(W, I)$; it facilitates the expectation over all possible transmitted constellation point sequences in the computation of $N(w, i)$. In this role the σ_x state diagram is a hidden Markov model with observables q. Any hidden Markov model (in particular one with the smallest possible number of states) that correctly produces the Markov process of equivalence class indices q can replace the σ_x state diagram without changing the final transfer function.

Let σ_q denote the states of a minimal hidden Markov model for the observables q. Let 2^{ν_q} be the number of states σ_q. The overall state diagram requires $2^{\nu_e + \nu_q}$ product states (σ_e, σ_q) and the transition labels are

$$P(s_q \to t_q) \sum_q \sum_{(e_i, e_o | s_e \to t_e)} P(q|s_q \to t_q) W^{d_E^2(q,e_o)} I^{d_H(e_i,0)}. \tag{9}$$

The simplest example of this reduction procedure yields the well-known result that the σ_e state diagram is always sufficient for convolutional codes with binary PAM. In this case $d_E^2(x, x \oplus e_o)$ depends only on the Hamming weight of e_o; there is only one equivalence class. As a second example, the regularity conditions of Zehavi [10] and Rouanne & Costello [11] essentially require every encoder state to be equally likely to produce every equivalence class. These conditions force the equivalence class sequence to be independent and identically distributed (IID), which allows $\nu_q = 0$.

These regularity conditions are not the only conditions for which the equivalence class sequence is IID. Furthermore, even when the sequence is not IID it is possible that $\nu_q < \nu_e$ avoiding the worst case requirement of $2^{2\nu_e}$ states. Often, equivalence class membership is determined by boolean operations on the constellation point labels. In this case, finding the minimal hidden Markov model is analogous to the well-known problem of finding a minimal range-equivalent convolutional encoder [13].

2.2 Computing the periodic transfer function

To construct the periodic state diagram from the minimal standard state diagram of the previous subsection, consider what happens on the standard state diagram in p steps. Suppose that there are m ways to transition from state (s_e, s_q) to state (t_e, t_q) in p steps. The periodic state diagram labels the transition $(s_e, s_q) \to (t_e, t_q)$ with the m-element sum of p-element products of branch labels along p-step paths from (s_e, s_q) to (t_e, t_q) in the standard state diagram. In this product, the variable W on the jth step is replaced by the variable W_j. Applying the substitution $W_j = e^{-\tilde{a}_j^2/4N_0}$ reduces symbolic manipulation. Note that the standard transition diagram and the periodic diagram both have the same number of states.

The computational procedures in [9] easily adapt to compute the periodic transfer function from our minimal periodic state transition diagram. Call the $N_q = 2^{\nu_q}$ product states where the $\sigma_e = 0$ the "correct states" and the remaining $N_e = 2^{\nu_e+\nu_q} - 2^{\nu_q}$ product states the "errored states". Define $\mathbf{A}$ to be the $N_e \times N_e$ matrix of labels for transitions from one errored state to another. Element $\mathbf{A}_{ts}$ specifies the transition label from state s to state t. Define $\mathbf{b}$ to be the $N_e \times N_q$ "tall" matrix for transitions from an errored state to a correct state and $\mathbf{c}$ to be the $N_q \times N_e$ "wide" matrix for transitions from a correct state to an errored state. Finally, let $\mathbf{d}$ be the $N_q \times N_q$ matrix for transitions from one correct state to another.

Computing the expectation $N(w_1, \ldots, w_p, i)$ over all possible sequences $\{q\}$ requires knowledge of the stationary distribution over the states σ_q. Let $\mathbf{q}$ be the p-element column vector of these stationary probabilities. Typically every element of $\mathbf{q}$ is $2^{-\nu_q}$. Finally, here is the equation for the periodic transfer function T_p:

$$T_p = \mathbf{1dq} + \sum_{j=0}^{\infty} \mathbf{1cA}^j\mathbf{bq} = \mathbf{1dq} + \mathbf{1c}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{bq}\,, \tag{10}$$

where $\mathbf{1}$ is the N_q element row-vector of ones and $\mathbf{I}$ is the $N_e \times N_e$ identity matrix. The derivative in (6) is

$$\frac{\partial T_p}{\partial I} = \mathbf{1d'q} + \mathbf{1c}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{b'q} + \mathbf{1c'}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{bq} + \mathbf{1c}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{A'}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{bq}\,, \tag{11}$$

where $\mathbf{d'}$ indicates the element-by-element partial derivative with respect to I.

As an example consider the rate-1/3, 8-PSK trellis code from [8] with polynomial 171 46 133 ($46 \rightarrow D^5 + D^2 + D$) and profile-superior [14] constellation labeling 0, 2, 3, 1, 5, 7, 6, 4 in octal going around the circle. The two equivalence classes for this labeled constellation are $\mathcal{C}_1 = \{000, 100, 001, 101\}$ and $\mathcal{C}_2 = \{010, 110, 011, 111\}$. Table 1 gives the Euclidean distance mappings for these two classes assuming $\mathcal{E}_x = 1$.

	Euclidean distances for each e_o							
$e_o \rightarrow$	000	100	001	010	101	110	011	111
q=1	0	0.5858	3.4142	0.5858	4	2	2	3.4142
q=2	0	3.4142	0.5858	0.5858	4	2	2	3.4142

Table 1: Mappings of symbol error e_o to squared Euclidean distance for the two equivalence classes comprising 8-PSK with labeling described above.

This trellis code does not satisfy the regularity conditions of [10] or [11] because the equivalence class produced during a state transition is completely determined by the encoder state σ_x. Despite this dependence, the equivalence-class random process is actually an IID sequence. The equivalence classes in this case are completely determined by the middle bit of the output symbol, we can view the equivalence class sequence as the output of a rate-1/1 convolutional code. The minimal range-equivalent encoder for this code requires no memory elements (i.e. $\nu_q = 0$). The labeling of (8) simplifies to

$$\frac{1}{2}\left(W^{d_E^2(q=1,e_o)} + W^{d_E^2(q=2,e_o)}\right) I^{d_H(e_i,0)}. \tag{12}$$

Converting to the periodic transfer function and computing (6) produces BER bounds in Figure 1. This figure shows the tightness of these bounds when compared to simulated BER for the six periodic symbol puncturing scenarios simulated in [8].

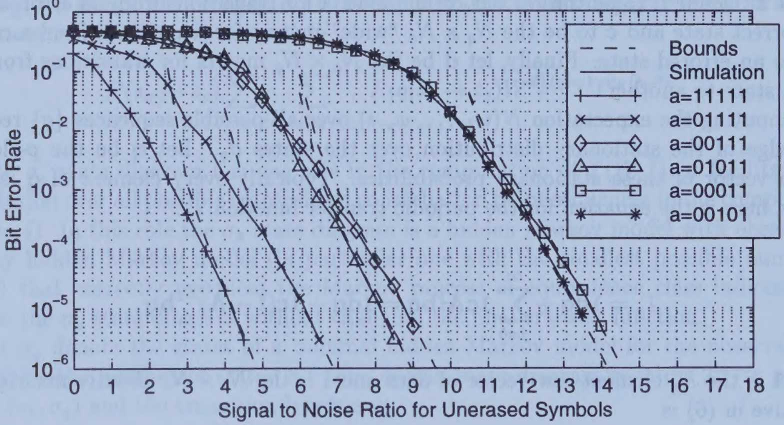


Figure 1: Transfer function BER bounds and simulations for a 64-state, rate-1/3, 8-PSK trellis code for six periodic erasure patterns on a period of five symbols.

3 Catastrophic Behavior Under Puncturing

A trellis encoder for the AWGN channel will perform much worse than expected given its nonzero free Euclidean distance if it is catastrophic in the sense that an infinite number of bit errors may result from a finite Euclidean distance error event. Any encoder that is minimal or systematic (or both) avoids this catastrophic behavior [13]. However, a periodic erasure pattern $\tilde{a}$ can cause catastrophic behavior despite a nonzero $\text{RED}^2(\tilde{a})$ even for an implementation that is minimal and systematic. The periodic erasures effectively produce a new, higher-rate encoder that may be neither systematic nor minimal.

A periodically punctured encoder is catastrophic if and only if there is a zero output Hamming weight path with nonzero input Hamming weight from any nonzero state to itself. Punctured symbols contribute no output Hamming weight. The following theorem relates catastrophic behavior in feedforward and feedback encoders.

Theorem 1 *If a feedback convolutional encoder is catastrophic under a fixed periodic erasure pattern, every range-equivalent feedforward convolutional encoder will also be catastrophic for that pattern.*

Proof: Since the pattern makes the feedback encoder catastrophic, the feedback encoder has a zero output Hamming weight path from some nonzero state to itself for this pattern.

Range equivalence ensures that the feedforward encoder also has this zero output Hamming weight path for this pattern, and the feedforward structure ensures that this path has nonzero input Hamming weight. Thus the feedforward encoder will be catastrophic under this erasure pattern. □

Converting an encoder from feedforward to feedback won't introduce new catastrophic behavior. Moreover, this conversion may actually prevent catastrophic behavior. As an example, consider code #7 from Table I of [15]. This rate-1/4 16-QAM feedforward trellis code has the following encoder polynomial:

$$G(D) = [D^5 + D^4 + D + 1 \quad D^6 + D^5 + D^4 + D^3 + D^2 + 1 \quad D^6 + D^4 + D^3 + D^2 + 1 \quad D^6 + D + 1] \, .$$

This code is catastrophic for the two erasure patterns $\tilde{a} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$ and $\tilde{a} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$. Converting this encoder to the range equivalent recursive encoder shown in Figure 2 produces an encoder that provides non-catastrophic behavior for every period-8 pattern with six or fewer erasures.

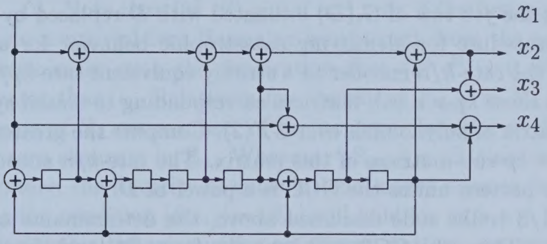


Figure 2: This recursive encoder is range-equivalent to feedforward encoder #7 in [15] and provides non-catastrophic behavior for every period-8 pattern with six or fewer erasures.

3.1 Determining catastrophic behavior in feedforward codes

One method for determining whether a feedforward encoder will display catastrophic behavior as a result of periodic erasures involves consideration of the encoder for a strictly equivalent multiple trellis code. Divsalar and Simon introduced multiple trellis codes in [16] to describe trellis codes that produce multiple constellation points with each state transition. Every rate-k/n trellis code has a strictly equivalent rate-kp/np multiple trellis code for which each state transition produces p constellation points. In the current context, each state transition produces the output symbols for one period of the periodic Gaussian channel. Strict equivalence to the original code implies that the multiple trellis code implements exactly the same mapping of input information sequences to output constellation-point sequences.

The multiple trellis code maintains the same number of memory elements as the original trellis code, but the exponent of the number of branches per state increases from k to kp. As an example, consider a rate-1/3, 8-PSK, $\nu = 3$ trellis code with minimal feedforward convolutional encoder

$$G_{1/3}(D) = \begin{bmatrix} 1 + D + D^3 & 1 + D^2 + D^3 & 1 + D + D^2 \end{bmatrix} . \tag{13}$$

For period $p = 3$, the strictly equivalent rate-3/9 multiple trellis code is

$$G_{3/9}(D) = \left[G_1(D) \quad G_2(D) \quad G_3(D)\right] \tag{14}$$

with the 3×3 sub-matrices $\left[G_1(D) \quad G_2(D) \quad G_3(D)\right]$ given in (15).

$$\left[\begin{bmatrix} 1+D & 1+D & 1 \\ 0 & D & D \\ D & 0 & D \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 1+D & 1+D & 1 \\ 0 & D & D \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1+D & 1+D & 1 \end{bmatrix}\right] \tag{15}$$

Each of the p sub-matrices $G_i(D)$ is a rate kp/n encoder fragment that produces the output of $G_{1/3}(D)$ shifted by i and downsampled (symbol-wise) by p. The original encoder can be recovered from any of these fragments. For $i \in \{1, 2, 3\}$ in our example,

$$G_{1/3}(D) = \sum_{j=1}^{p} D^{i-j} G_i^{(j)}(D^3), \tag{16}$$

where $G_i^{(j)}(D^3)$ is the jth row of $G_i(D)$ evaluated with D replaced by D^3.

Here is the procedure for identifying catastrophic behavior for a specified erasure pattern: Expand the rate-k/n encoder to a strictly equivalent rate-kp/np multiple trellis encoder. Remove those $kp \times n$ sub-matrices corresponding to erased symbols, producing a $kp \times n(p-q)$ matrix of polynomials over $GF(2)$. Compute the greatest common divisor (GCD) of all $kp \times kp$ sub-matrices of this matrix. The rate-k/n encoder is catastrophic under the erasure pattern unless the GCD is a power of D.

For the rate-1/3 trellis code discussed above, the determinants of G_1, G_2, and G_3 are D^2, D, and 1. Thus, the GCD will have the form D^j (and the overall system will be non-catastrophic) even when two of three symbols are erased. As a second example, consider the popular rate-1/2, $\nu = 6$ convolutional code

$$G(D) = \left[1 + D + D^3 + D^4 + D^6 \quad 1 + D^3 + D^4 + D^5 + D^6\,.\right] \tag{17}$$

The strictly equivalent rate-2/4 multiple trellis code is

$$G_{2/4}(D) = \left[\begin{array}{cc|cc} 1+D^2+D^3 & 1+D^2+D^3 & 1+D & D+D^2 \\ D+D^2 & D^2+D^3 & 1+D^2+D^3 & 1+D^2+D^3 \end{array}\right] \tag{18}$$

When every other symbol is periodically punctured, the remaining rate-2/2 sub-code either has determinant $D + D^4 + D^5 + D^6$ or $1 + D^3 + D^4 + D^5$. Since these determinants are not of the form D^j this rate-1/2 code becomes catastrophic under periodic puncturing of every other symbol.

3.2 General algorithm for identifying catastrophic behavior

Lapidoth's algorithm for computing BER in noiseless periodic erasure channels, Theorem 1 in [4], provides the foundation for a computational method for determining catastrophic behavior. Define a $p - q$ erasure pattern as a pattern $\tilde{a}$ with $\tilde{a}_i \in \{0, 1\}$ and q zeros. To obtain the procedure for identifying catastrophic behavior, we first simplify Lapidoth's algorithm and then generalize the resulting procedure.

Our simplified version of Lapidoth's algorithm simply determines whether BER is zero, i.e. whether $\langle \tilde{a}^2, \tilde{d}^2 \rangle \neq 0$ for every $\tilde{d}^2$ and a specified erasure pattern $\tilde{a}$ on an otherwise noiseless channel. Following [4], consider two matrices, A_0 and A_1, that provide partial information about the σ_e state diagram for $\tilde{a}_i = 0$ (erased symbol) and $\tilde{a}_i = 1$ (no erasure), respectively. These matrices have only zeros and ones as entries. For both matrices, $A(s,t) = 1$ if and only if a transition from s to t is possible and cannot be distinguished from the zero-state self loop by the (possibly erased) encoder output. For an erased symbol, $A_0(s,t) = 1$ if there is a branch connecting state s to state t in a single step. Otherwise, $A_0(s,t) = 0$. For an unerased symbol, $A_1(s,t) = 1$ if there is a branch connecting state s to state t in a single step and the convolutional encoder output for this branch is all zeros. Otherwise, $A_1(s,t) = 0$. To consider the state transitions possible in p steps corresponding to one period of the erasure pattern, let

$$A_{\tilde{a}} = \prod_{i=1}^{p} A_{\tilde{a}_i} \, . \tag{19}$$

Define S_m to be the 2^ν-element row vector of integers where the $j+1$ entry of S_m is the number of distinct zero Hamming weight paths from the zero state to state j after mp state transitions. S_m can be computed through the recursion $S_{m+1} = S_m A_{\tilde{a}}$ with $S_0 = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \end{bmatrix}$. Note that $\langle \tilde{a}^2, \tilde{d^2} \rangle \neq 0$ if and only if the first entry of S_m is 1 for all m, i.e. there is no nontrivial zero Hamming weight path from the zero state to itself.

Computation may cease with the declaration that $\langle \tilde{a}^2, \tilde{d^2} \rangle = 0$ whenever the first entry of S_m is greater than 1. Relatively few iterations $S_{m+1} = S_m A_{\tilde{a}}$ are necessary. The nonzero elements of S_m are a subset of the nonzero elements of S_{m+1}. Let $|S_m|$ be the number of nonzero elements in S_m. Whenever $|S_{m+1}| = |S_m|$ no more iterations are necessary. Furthermore, the linearity of convolutional codes ensures that $|S_m|$ is always a power of 2. Thus the condition $|S_{m+1}| = |S_m|$ will hold for $m > \nu + 1$.

To identify catastrophic behavior under an erasure pattern, we generalize this technique to identify zero Hamming weight paths from any state to itself. Define S_0^j to be the 2^ν-element row vector that is all zeros except a one in the jth position, and let $S_{m+1}^j = S_m^j A_{\tilde{a}}$. A feedforward encoder is non-catastrophic under the erasure pattern $\tilde{a}$ if and only if the jth entry of S_m^j is zero for all $j \in \{1, 2^\nu - 1\}$ and all $m > 0$. For feedback encoders, catastrophic behavior occurs only when one of the zero output Hamming weight loops has an associated input Hamming weight that is nonzero.

4 Conclusions

Transfer function techniques provide asymptotically tight bounds for the combined effects of AWGN and a specified periodic fading pattern. Unlike almost all good trellis codes for AWGN channels, trellis codes that perform well on channels with periodic fading often are not regular in the sense of Zehavi [10] or Rouanne and Costello [11]. We presented a procedure to efficiently compute the transfer function regardless of whether a trellis code is regular.

Only non-minimal encoders display catastrophic behavior on the AWGN channel. However, periodic erasures can make even minimal encoders display catastrophic behavior. We presented two procedures for identifying which erasure patterns cause a trellis code to fail in this way. In terms of this catastrophic behavior, feedback encoders are preferable. They may avoid catastrophic behavior by assigning zero input Hamming weight to the zero output Hamming weight loop responsible for the catastrophic behavior in a range-equivalent feedforward implementation.

References

[1] J. Hagenauer. Rate-Compatible Punctured Convolutional Codes (RCPC Codes) and their Applications. *IEEE Trans. on Comm.*, 36(4):389–400, April 1988.

[2] R. D. Wesel and J. M. Cioffi. Joint trellis code and interleaver design. In *Globecom '97*, Nov. 1997.

[3] Y. S. Leung, S. G. Wilson, and J. W. Ketchum. Multifrequency Trellis Coding with Low Delay for Fading Channels. *IEEE Trans. on Comm.*, 41(10):1450–1459, Oct. 1993.

[4] A. Lapidoth. The performance of convolutional codes on the block erasure channel using various finite interleaving techniques. *IEEE Trans. on Info. Th.*, 40(5):1459–1473, Sept. 1994.

[5] R. D. Wesel and J. M. Cioffi. Fundamentals of coding for broadcast OFDM. In *Twenty-Ninth Asilomar Conference on Signals, Systems, and Computers*, Nov. 1995.

[6] R. D. Wesel. *Trellis code design for correlated fading and achievable rates for Tomlinson Harashima precoding.* PhD thesis, Stanford University, Aug. 1996.

[7] R. Knopp and P. A. Humblet. Maximizing diversity on block-fading channels. In *Proc. of ICC-97*, pages 647–651, June 1997.

[8] R. D. Wesel, X. Liu, and W. Shi. Periodic symbol punturing of trellis codes. In *Thirty-First Asilomar Conference on Signals, Systems, and Computers*, Nov. 1997. invited.

[9] E. Biglieri. High-Level Modulation and Coding for Nonlinear Satellite Channels. *IEEE Trans. on Comm.*, 32(5):616–626, May 1984.

[10] E. Zehavi and J. K. Wolf. On the performance evaluation of trellis codes. *IEEE Trans. on Info. Th.*, 33(2):196–202, March 1987.

[11] M. Rouanne and D. J. Costello, Jr. An algorithm for computing the distance spectrum of trellis codes. *J. of Sel. Areas in Comm.*, 7(6):929–940, August 1989.

[12] S. Benedetto, M. Mondin, and G. Montorsi. Performance evaluation of Trellis Coded Modulation Schemes. *Proc. IEEE*, 82(6):833–855, June 1994.

[13] G. D. Forney, Jr. Convolutional Codes I: Algebraic Structure. *IEEE Trans. on Inform. Theory*, 16(6):720–738, Nov. 1970.

[14] R. D. Wesel and J. M. Cioffi. Towards optimality in constellation labeling. In *Globecom '97*, Nov. 1997.

[15] R. D. Wesel and J. M. Cioffi. Trellis Code Design for Periodic Interleavers. *IEEE Comm. Let.*, submitted.

[16] D. Divsalar and M. K. Simon. Multiple Trellis Coded Modulation (MTCM). In *IEEE Global Telecomm. Conf. Rec.*, pages 30.8.1–30.8.7, Houston,TX, Dec. 1986.

An Interleaving Technique for Error Correction in 2-D Bar Coding

George F. Elmasry Yun Q. Shi
georgelisa@aol.com shi@tesla.njit.edu
New Jersey Center for Multimedia Research
Department of Electrical and Computer Engineering
New Jersey Institute of Technology, Newark, NJ 07102-1982

Abstract

In two-dimensional bar coding, the error correction decoder has access to the entire encoded data set at once, in the form of a 2-D array of symbols. In this 2-D array, errors occur in bursts that stretch across the horizontal and vertical axes, corrupting a cluster of symbols in the array. This paper introduces an interleaving technique that fully interleaves the encoded sequence of symbols for this type of error. We will compare the performance of this technique with the performance of block interleaving when applied to 2-D bar coding.

1 Introduction

2-D bar codes are information storage media in which the source information is stored as a binary stream on a printed label. Examples of 2-D bar code applications are numerous. In hospitals, bracelets can carry a patient's entire medical history. In the automotive assembly process, a label on a part can carry a unique identification number and other pertinent information applicable to production, tracking, and statistical process control [1]. In the shipping industry, the 2-D bar code label can be used as an alternative for EDI (Electronic Data Interchange). That is, the label is a portable data file that accompanies the package [2]. The size of the label is decided by the number of elements to be encoded and the size of the element [3]. The *element* is the smallest area that can be printed and read using available technology. To maximize the information contents of the label, the size of an element (one element carries one bit of encoded data) should be just large enough for the decoding device (scanner or camera) to be able to decide whether this element represents a binary 1 or a binary 0. The resolution of the print mechanism can be expressed in terms of the number of square pixels contained in the element. Figure 1-a shows the magnification of 2 print patterns for dark elements used in the UPS MaxiCode [2]. Each square dark area is one pixel, and the hexagon shape is one element. The UPS MaxiCode carries up to a hundred information characters in a one-inch by one-inch area. The reader may refer to works such as [1], [2], [3], and [4] for more details about 2-D bar coding.

There are some similarities and some differences between 2-D bar coding and CD. As with the audio CD recording mechanism [5], the label printer in 2-D bar coding performs two successive coding processes: (a) error correction and (b) writing (recording). The source alphabet size of the encoded stream depends on the specific 2-D bar coding application. In the automotive assembly process, one application of 2-D bar coding is to encode hexadecimal computer instructions to be read by robots to implement particular operations [1]. This application requires 4 bits per symbol. Designers of error correction codes for such an application use Reed-Solomon (RS) codes from $GF(2^4)$. In the shipping industry, the need to encode numerical and English characters, in addition to some punctuation, requires using 6 bits per symbol and hence RS codes are constructed from $GF(2^6)$ [2]. The specifications published in [4] explain encoding 7-bit ASCII symbols and 8-bit extended ASCII symbols where RS codes are constructed from $GF(2^7)$ and $GF(2^8)$ respectively. The writing code translates the sequence of symbols resulting from the error correction code to a 2-D array of bits, and the label is printed accordingly. Binary 1 and 0 can be expressed as dark and white elements or vice versa. To achieve better error protection capability using RS codes, the writing code attempts to cluster the elements representing one symbol in one area. Figure 1-b shows how writing multiple lines of elements in parallel achieves this for 4-bit and 6-bit symbols. To understand why this clustering is needed, consider capturing the image of the label by a scanner or a camera from a warped surface or partially damaged label: the number of symbols containing errors is less with clustering than if the bit stream is written directly to the label, line by line.

There are some major differences between 2-D bar coding and CD which should decide the best interleaving technique needed. With CD, the decoder does not have access to all the data at once, and hence interleaving is done to one block of data at a time. With 2-D bar codes, the decoder has access to all the data at once in the form of a 2-D array. This is because the image of the label is captured and processed, resulting in a 2-D array of symbols available for the error correction decoder. The second difference is that damage to the storage media in a CD is likely to occur from a scratch that stretches across the spiral pattern created in the recording of the CD, making block interleaving a suitable choice. With 2-D bar codes, the damage to the storage media can result in a loss of information in one area of the label.

This paper addresses the following question: For 2-D bar codes, knowing that the decoding device has access to all the data at once in the form of a 2-D array, and knowing that the error bursts stretch across both vertical and horizontal axes, what is the best interleaving strategy for the error correction code to follow? In other words, starting from well known block interleaving, which is a 1-D interleaving technique that spreads *consecutive* errors across the interleaved block, can we come up with a 2-D interleaving technique that spreads *neighboring* errors across the interleaved sequence? "Neighboring" means that errors are adjacent across both vertical and horizontal axes (that is, a subset of symbols in the array is damaged). In the next section we will show how block interleaving can be achieved by visualizing a 1-D interleaving array and utilizing less memory space than that needed by the conventionally used $n * n$ interleaving array. In section 3 we will use this analysis to develop a 2-D interleaving technique for 2-D arrays that is equivalent to block interleaving for 1-D arrays. In section 4 we will compare 2-D array interleaving with block interleaving when used in 2-D bar codes.

2 Block Interleaving as a 1-D Interleaving Technique

Suppose we have a block of length n^2 symbols, which is a 1-D array, and we want to apply block interleaving to it. Maximum interleaving can be obtained by using a 2-D memory array of size $n * n$ symbols [6] [7] [8]. Symbols are written to the memory array column by column and read row by row (or vice versa). In this case, the *interleaving degree* known as λ, is equal to n. If the RS code used is capable of correcting a single burst error of length l symbols or less in a block of length n, the interleaved code will correct any single burst of length λl or less. Figure 2 demonstrates this technique when $n = 4$ and $l = 1$; that is, the sequence to be interleaved is a 1-D array of 16 symbols. Figure 2-a shows the original sequence, Figure 2-b shows the interleaving $4 * 4$ array used, and Figure 2-c shows the resulting interleaved sequence. If 4 consecutive symbols are in error as shown by the shaded areas in Figure 2-c, the de-interleaved array will have these errors spread out as shown in Figure 2-d, where each block of 4-symbols contains only one symbol in error. If we use an RS code capable of correcting one symbol in a four-symbol block, we will succeed to recover the original data. Without interleaving, the same RS code will fail to recover the original data. Notice that if we divide the de-interleaved array into 4 equal segments, any 4 consecutive symbols in the interleaved array are spread in the de-interleaved array such that one symbol falls in each segment, which ensures the desired correction capability.

To apply the same concept to a 2-D array of symbols, one may think of utilizing a 3-D or 4-D interleaving array (which is hard to visualize) and find an algorithm to read from and write to this interleaving array such that maximum interleaving is achieved. This algorithm may require a large memory space and extensive computation. Another approach is to interleave 1-D arrays using 1-D interleaving arrays and utilize as little memory space and computation power as possible, and adopt this method for 2-D arrays. Figure 3 shows how this 1-D interleaving is done for the case of $n = 4$, explained above. Figure 3-a shows a 1-D interleaving array of 5 symbols. We will name this interleaving array I. In I, we will call the location at the center q and the end locations e_1 and e_2. Figure 3-b shows an empty array of 16 symbols with 8 locations q_1 to q_8 marked at the positions shown. We will name this array R. Imagine that we write the first 2 symbols of the original sequence to e_1 and e_2 in I, locate I such that the center q is at location q_1, and write the 2 symbols from e_1 and e_2 to their corresponding locations in array R. This will be the first and fifth locations in R. Afterwards, we write the next two symbols from the original array to e_1 and e_2, move I such that the center q is at location q_2, and write from e_1 and e_2 to the corresponding locations in R. This will be the ninth and thirteenth locations. If we repeat these steps until we cover the eight marked locations in R, we will end up with symbols written to R in the same sequence as in Figure 2-c. One can see that, in general, the length of I is $1 + \lambda$, where λ is the interleaving degree. Also, the number of locations that need to be marked for q to follow is $n^2/2$.

To apply this method practically, aside from the original array and the interleaved array, we will need an array of size $n^2/2$ memory locations to contain the indexing of the marked locations $q_i's$ with respect to R. We will name this array the q array. For the example of $n = 4$ explained above, the q array will contain the indexes 2, 10, 3, 11, 4, 12, 5, and 13 respectively. The short program (written in C) in Appendix A shows how only 8 memory locations are utilized using these 8 indexes instead of the 16 memory locations used by block interleaving.

One may argue about the way in which we can obtain the locations q_1 to q_8. The next section will show how a simple iterative algorithm can decide the locations q_i's for different 2-D array sizes. We will show the convenience of applying this concept for 2-D interleaving, since it does not require utilizing a complicated interleaving array or a large memory space, and it can be implemented in real time.

3 Interleaving for 2-D Arrays

In this section, interleaving for 2-D arrays is explained for $n * n$ arrays where n is a power of 2. The interleaving technique visualizes a 2-D interleaving array, which we will refer to as I. The dimensionality of I is $p * p$ symbols with the four corner symbols e_1, e_2, e_3, and e_4 arranged as shown in Figure 4. p is decided by n through the following relation:

$$p = n/2 + 1, \tag{1}$$

and the center of I is **in** from the 4 corners by q' symbols, where q' is given by

$$q' = p - median(p). \tag{2}$$

The locations of q denoted as q_i's form another 2-D array, which will be referred to as the q array. The q array is a subset of the 2-D array referred to as R (which will carry the interleaved symbols). The q array has a dimension of $n/2 * n/2$ and is located q' symbols **in** from the edges of the R array. The application of this algorithm is iterative; that is, to interleave an $n * n$ array, we obtain the interleaved $n/2 * n/2$ array first and use its sequencing as the q array sequencing. The following cases explain this interleaving technique. The reader can verify from these cases that the algorithm interleaves across the horizontal axis with a degree $\lambda_x \geq n/4$ and across the vertical axis with a degree $\lambda_y \geq n/4$ such that the total interleaving $\lambda = \lambda_x + \lambda_y$ is lower-bounded by $n/2$; that is, $\lambda \geq n/2$.

For n = 2: In this case equation (1) gives $p = 2$, and the R array is a copy of the I array itself as shown in Figure 5. Note how the interleaving degree $\lambda \geq 1$; between symbols 1 and 2, $\lambda_x = 1$ and $\lambda_y = 1$; between symbols 2 and 3, $\lambda_x = 0$ and $\lambda_y = 1$; and between symbols 3 and 4, $\lambda_x = 1$ and $\lambda_y = 1$.

For n = 4: In this case equation (1) gives $p = 3$ and equation (2) gives $q' = 1$. The original sequence is shown in Figure 6-a. The interleaving array I is shown in Figure 6-b, and the q array is located 1 symbol inside the R array as shown in Figure 6-c. The sequencing of the q array is the same as the interleaved array obtained previously for $n = 2$ (compare Figure 6-c with Figure 5-c). Applying the following three steps 4 times for $i = 1, 2, 3$, and 4, the interleaved array shown in Figure 6-d is obtained:

1. e_1, e_2, e_3, and e_4 read from the original array shown in Figure 6-a.
2. The I array center q is located at q_i.
3. e_1, e_2, e_3, and e_4 write to the corresponding locations in the R array.

One can see that $\lambda_x \geq 1$, $\lambda_y \geq 1$, and $\lambda \geq 2$. Notice the resemblance to block interleaving in the sense that if we divide the de-interleaved sequence (which is the same sequence in Figure 6-a) into 4 equal segments (code blocks), any cluster of $2 * 2$ symbols in the interleaved array is spread in the de-interleaved sequence such that one symbol falls in each segment, which ensures the desired correction capability. Similarly, if we divide the de-interleaved sequence into 8 equal segments, any cluster of $2 * 4$ symbols (1/2 of the label area) in the interleaved array is spread in the de-interleaved sequence such that one symbol falls in each segment. To apply this algorithm in real time (as explained for the 1-D case in the previous section) using the above three steps, we will need two memory locations for each entry in the q array for the x and y coordinates. That is, we will need an array of size $n^2/2 = 8$, as in the 1-D case. This array will carry the coordinates (1,1; 2,2; 2,1; and 1,2). The short program in Appendix B shows how this interleaving can be performed.

For n = 8: In this case $p = 5$ and $q' = 2$. The interleaving array I is shown in Figure 7-a, and the q array is located 2 symbols inside the R array as shown in Figure 7-b. The sequencing of the q array is the same as the interleaved array obtained previously for $n = 4$ (compare Figure 7-b with Figure 6-d). Applying the same three steps above 16 times for $i = 1, 2, ...$ and 16, the interleaved 2-D array shown in Figure 7-c is obtained. This array can present a 2-D bar code that contains 64 encoded characters, which is a possible practical application. In this case $\lambda_x \geq 2$, $\lambda_y \geq 2$, and $\lambda \geq 4$. $n^2/2 = 32$ memory locations are needed for interleaving. The resemblance to block interleaving can be observed if we divide the de-interleaved sequence into 16 equal segments. Any cluster of $4 * 4$ symbols in the interleaved array is spread in the de-interleaved sequence such that one symbol falls into each segment, which ensures the desired correction capability. Also, if we divide the de-interleaved sequence into 4 equal segments, any cluster of $2 * 2$ symbols in the interleaved array is spread in the de-interleaved sequence such that one symbol falls in each segment.

4 Comparison

Consider the case when $n = 4$, and 4 neighboring symbols are in error in the interleaved array as shown in Figure 8-a. When the de-interleaved sequence is obtained, the error will be spread across it as shown in Figure 8-b. If the RS code used is capable of correcting one error in a four-symbol block, it will succeed to recover the original data. Without interleaving, the same RS code would fail. As explained in the previous section, this correction capability applies for any single cluster of $2 * 2$ symbols in error. When considering the case of $n = 8$, we showed that, for the same correction capability of one error in four symbols, this interleaving technique will succeed to correct any single cluster of $4 * 4$ symbols in error. In the $n = 8$ case, if the RS code used is capable of correcting 1 error in 16 symbols, the original information will be recovered if any single cluster of $2 * 2$ symbols is in error.

Let us compare the performance of RS codes when the error correction code in 2-D bar codes applies the proposed 2-D interleaving technique versus block interleaving. In [2], block interleaving is used with two symbol writing strategies. With the first strategy, called the "boustrophedonic" pattern (shown in Figures 9-a and 9-b), symbols are written alternately in opposite directions, from left to right and right to left after applying block interleaving. With the second strategy, called the spiral data pattern (shown in Figures 9-c and 9-d), symbols are written in a rough spiral moving outward (much like a CD recording) after applying block interleaving. Figures 9-b and 9-d show the resulting de-interleaved sequence for both strategies. With the same error correction code and the same error pattern of a $2 * 2$ cluster of symbols, one can see that the error correction code will fail to recover the data using both strategies. One may think that for large arrays, the spiral data pattern with block interleaving can result in optimum interleaving as the spiral moves outward. As can be verified from Figure 9-e, where block interleaving is applied with the spiral pattern for the $n = 8$ case, this is not so. Consider any round of the spiral. There is no guarantee that symbols in the next or previous round (which can fall in the same damaged area) will not share the same segment (code block) in the de-interleaved sequence, which would result in a block coding error. For 2-D bar-code designers, relying on the writing code to help in symbol interleaving is needed, because an interleaving technique equivalent to block interleaving for 2-D cases is not currently available in error-correction coding references.

5 Conclusions

This paper addresses interleaving for 2-D bar coding applications. The interleaving technique introduced optimizes interleaving across both vertical and horizontal axes such that error correction is possible if the label is partly damaged. This shows that for 2-D bar code designers, 2-D interleaving algorithms in error control coding references are needed. The comparison in section 4 demonstrates the advantage of using 2-D interleaving over block interleaving in 2-D bar codes.

References

[1] John Piatek, “Automotive Industry Recommends 2-D Standards,” *Automatic I.D. News*, August 1994, pp. 54-56.

[2] MaxiCode Briefing Document, UPS Inc., 1996, http://www.maxicode.com.

[3] T. Pavlidis, J. Swartz, and Y. P. Wang, “Information Encoding with Two-Dimensional Bar-Codes,” IEEE *Computer* magazine, June 1992, pp. 18-28.

[4] Uniform Symbology Specification Code One, AIM Technology Group, 1994, TSC040.

[5] Kees A. Schouhamer Immink, “Reed-Solomon Codes and the Compact Disc,” in *Reed-Solomon Codes and Their Applications,* S. B. Wicker and V. K. Bhargava (Eds.), New York: IEEE press, 1994, pp. 41-59.

[6] S. B. Wicker, *Error Control Systems for Digital Communication and Storage.* Englewood Cliffs, NJ: Prentice-Hall, Inc., 1995.

[7] Shu Lin and Daniel J. Costello, Jr., *Error Control Coding: Fundamentals and Applications.* Englewood Cliffs, NJ: Prentice-Hall, Inc., 1983.

[8] S. G. Wilson, *Digital Modulation and Coding.* Upper Saddle River, NJ: Prentice-Hall, Inc., 1996.

APPENDIX A

```
/* 1-D interleaving for n = 4 */
#include <stdio.h>
main(){
int i;
/* Original Array */
int Org[16] = {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16};
/* Resulting Array */
int R[16];
/* indexing Array */
int q[8] = { 2, 10, 3, 11, 4, 12, 5, 13 };
   for(i = 0; i < 8; i++){
      R[q[i] - 2] = Org[2*i];
      R[q[i] + 2] = Org[2*i + 1];
   }
   for(i = 0; i < 16; i++) printf("%d ", R[i]);
}
```

APPENDIX B

```
/* 2-D interleaving for n = 4 */
#include <stdio.h>
main(){
int i, j = 0;
/* Original Array */
int Org[4][4]={1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16};
/* Resulting Array */
int R[4][4];
/* indexing Array */
int q[2][4] = { 1, 2, 2, 1,
                1, 2, 1, 2 };
   for(i = 0; i < 4; i++){
      R[q[0][i]-1][q[1][i]-1] = Org[i][j++];
      R[q[0][i]+1][q[1][i]+1] = Org[i][j++];
      R[q[0][i]+1][q[1][i]-1] = Org[i][j++];
      R[q[0][i]-1][q[1][i]+1] = Org[i][j];
      j = j - 3;
   }
   for(i = 0; i < 4; i++){
      for(j = 0; j < 4; j++){
         printf("%d ", R[j][i]);
      }
   }
}
```

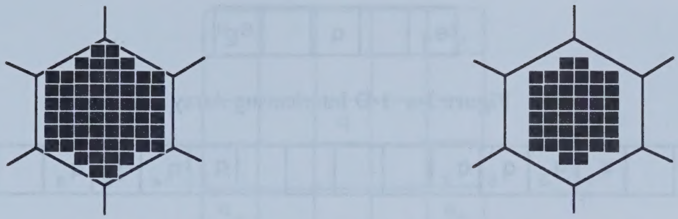
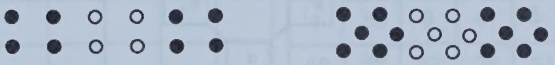


Figure 1-a Two Print Pattern for Dark Elements Used in the UPS MaxiCode



4-bit word

6-bit word

Figure 1-b Grouping the Elements of One Symbol in One Area for Burst Error Correction

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Figure 2-a Data Before Interleaving

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

Figure 2-b Two-Dimensional 4_*4 Array Used For Interleaving

1	5	9	13	2	6	10	14	3	7	11	15	4	8	12	16

Figure 2-c Data After Interleaving

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Figure 2-d Data After De-interleaving

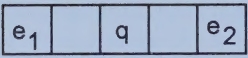


Figure 3-a 1-D Interleaving Array *I*

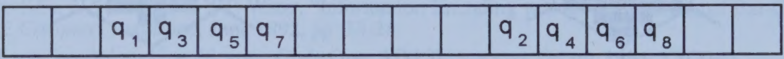

q1
q3
q5
q7
q2
q4
q6
q8

Figure 3-b Positions of the Interleaving Array Center q_i in the Interleaved Array *R*

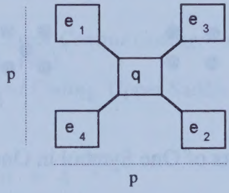
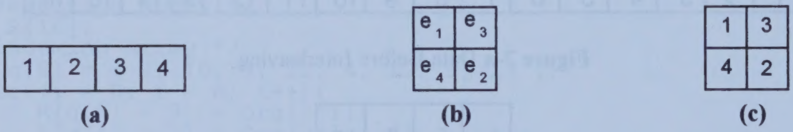

e1
e3
p
q
e4
e2
p

Figure 4 The Interleaving 2-D Array *I*


1 2 3 4
e1 e3
e4 e2
1 3
4 2

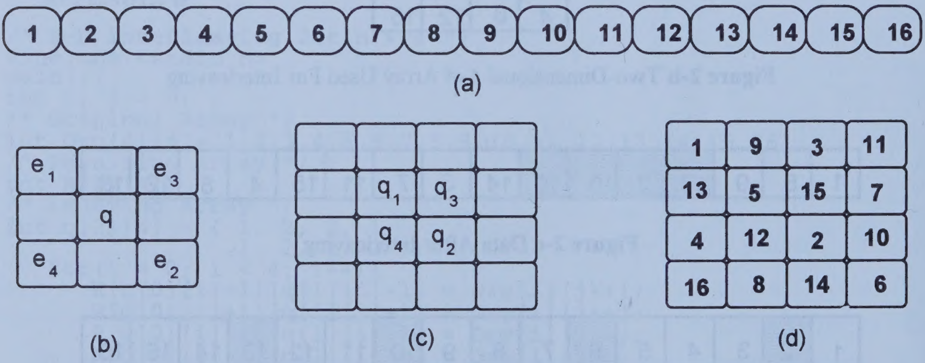
(a)

(b)

(c)

Figure 5 Interleaving for n = 2

(a) The Original Sequence (b) The Interleaving Array *I*. (c) The Interleaved Array


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
e1 e3
q
e4 e2
q1 q3
q4 q2
1 9 3 11
13 5 15 7
4 12 2 10
16 8 14 6

(a)

(b)

(c)

(d)

Figure 6 Interleaving For *n*=4

(a) The Original Sequence

(b) The Interleaving Array *I*

(c) Position of the *q* Array

(d) The Interleaved Array in 2-D (label layout)

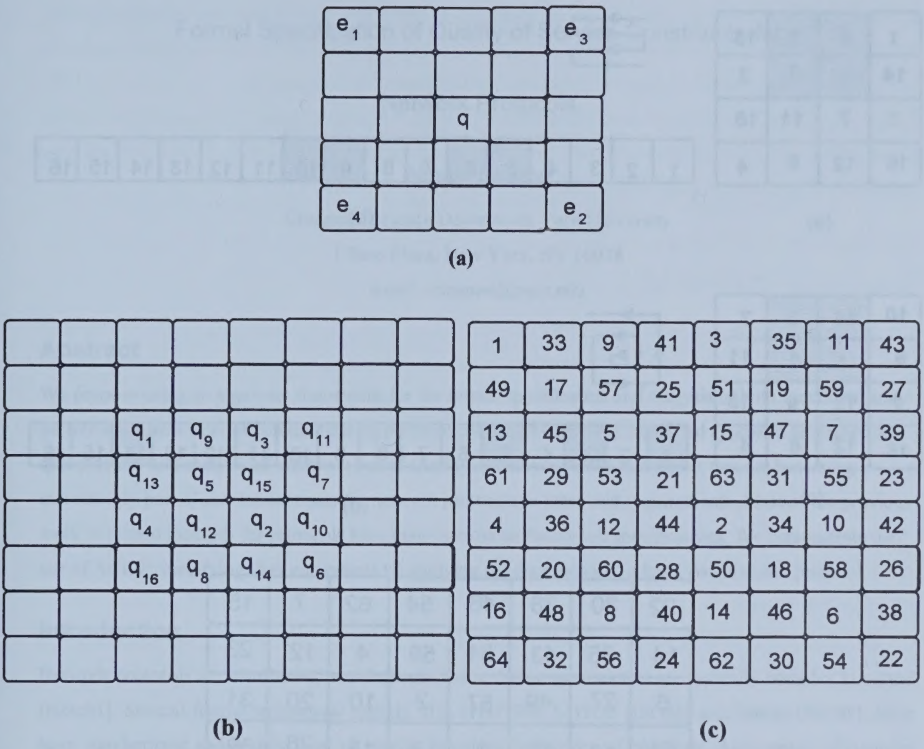

e1 e3
q
e4 e2
(a)
q1 q9 q3 q11
q13 q5 q15 q7
q4 q12 q2 q10
q16 q8 q14 q6
(b)
1 33 9 41 3 35 11 43
49 17 57 25 51 19 59 27
13 45 5 37 15 47 7 39
61 29 53 21 63 31 55 23
4 36 12 44 2 34 10 42
52 20 60 28 50 18 58 26
16 48 8 40 14 46 6 38
64 32 56 24 62 30 54 22
(c)

Figure 7 Interleaving for n=8
(a) The Interleaving Array I (b) Position of the q Array (c) Interleaved 2-D Array

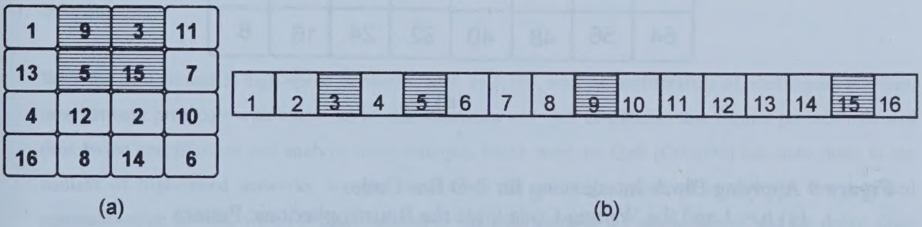

1 9 3 11
13 5 15 7
4 12 2 10
16 8 14 6
(a)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
(b)

Figure 8 Error Correction Capability for Interleaved 2-D Array
(a) Error Pattern in the Interleaved Array (label layout)
(b) Error Pattern in the De-interleaved Sequence

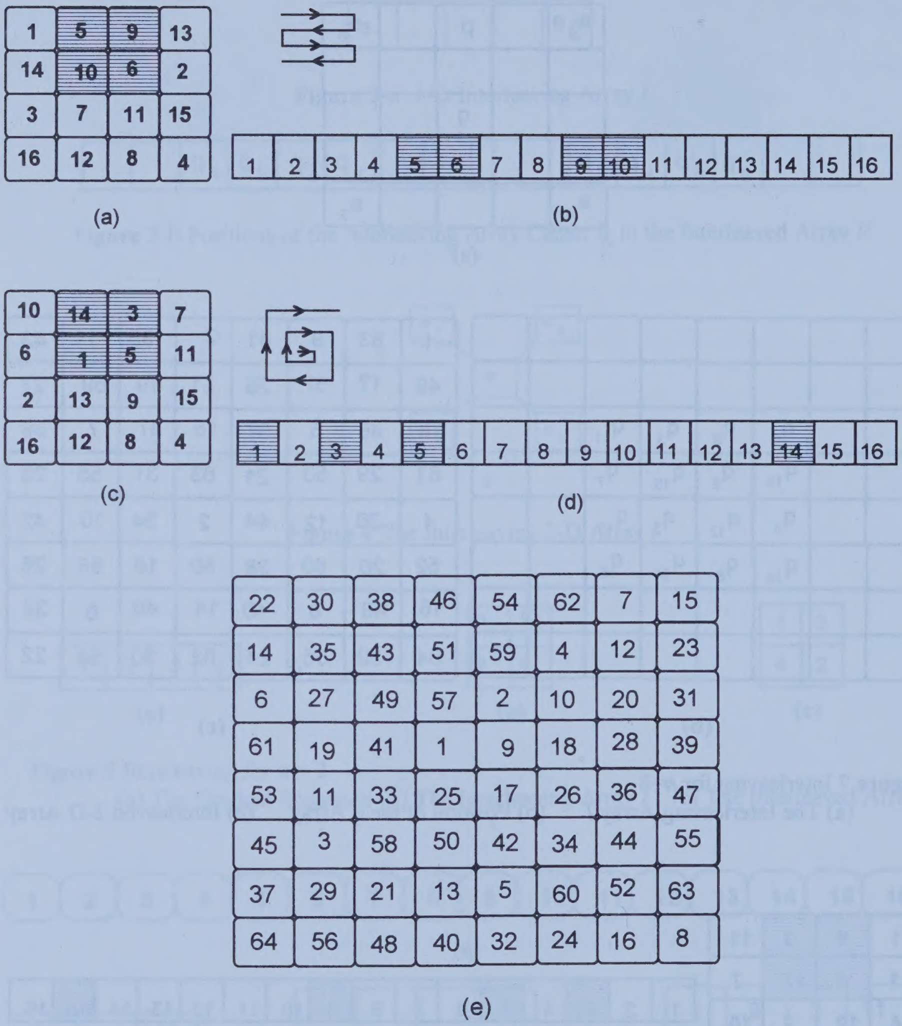

1 5 9 13
14 10 6 2
3 7 11 15
16 12 8 4
(a)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
(b)
10 14 3 7
6 1 5 11
2 13 9 15
16 12 8 4
(c)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
(d)
22 30 38 46 54 62 7 15
14 35 43 51 59 4 12 23
6 27 49 57 2 10 20 31
61 19 41 1 9 18 28 39
53 11 33 25 17 26 36 47
45 3 58 50 42 34 44 55
37 29 21 13 5 60 52 63
64 56 48 40 32 24 16 8
(e)

Figure 9 Applying Block Interleaving for 2-D Bar Codes
(a) n = 4 and the Writing Code Uses the Boustrophedonic Pattern
(b) The De-interleaved Symbols
(c) n = 4 and the Writing Code Uses the Spiral Data Pattern
(d) The De-interleaved Symbols
(e) n = 8 and the Writing Code Uses the Spiral Data Pattern

Formal Specification of Quality of Service Constraints for Network Protocols

Nihal Nounou

Computer Science Department, Pace University
1 Pace Plaza, New York, NY 10038
email: nnounou@pace.edu

Abstract

We propose using an algebraic framework for the formal specification and analysis of both functional and performance behaviors of communicating protocol entities (ACE). This integrated approach facilitates the specification and analysis of various types of Quality of Service (QoS) constraints that are becoming increasingly part of service contracts for network protocols to meet user requirements. Most of the previous work in formal methods for protocols have concentrated on functional specifications. We demonstrate the use of ACE in specifying QoS constraints by applying the framework to a basic data transfer protocol.

Introduction

Network protocols are distributed, non-deterministic real-time systems that are typically complex in nature [Holz91]. Several formal techniques, namely SDL [ITUT93], LOTOS [ISO88] and Estelle [ISO89], have been standardized and widely used to specify functional behaviors of protocols. Advantages of rigorous formal specifications include avoiding inconsistencies in the requirements, serving as a basis for deriving correct system designs, establishing the system's correctness through testing and formal verification, and providing unambiguous documentation of system requirements. Some efforts have been lately underway to address the real-time and probabilistic features and constraints of protocols in addition to their functional behavior.

With the deployment of high-speed networks, such as ATM, and the proliferation of multimedia services, new network protocols with a new set of characteristics and QoS constraints that require probabilistic and time-based specification and analysis have emerged. Much work on QoS [Camp94] has been done in the context of high-speed networks in order to provide for some guarantee of quality for the provided communication service, which is characterized by the bandwidth of the media stream and the delay, jitter and loss rate of the network. While in traditional systems the notion of system correctness is defined in terms of the functional correct ordering of externally observable events, new protocols are deemed correct if they satisfy both functional correctness requirements and quantitative QoS requirements.

In this paper, we investigate the use of an algebraic framework for the formal specification of both functional and performance behaviors of communicating protocol entities (ACE). This integrated approach facilitates the specification and analysis of various types of Quality of Service (QoS).

We present the ACE framework next. Formal proofs are not completely given in this paper, but can be found in [Noun86]. We then describe how ACE can be used to formally specify QoS constraints of network protocols with a basic data transfer protocol as an example.

Specification of Behaviors

The behavior of each communicating entity is modeled by a set of algebraic equations. The special behavior "$" indicates termination. The constants in the algebra are send, receive, or rendezvous events, denoted by lower case strings and preceded by a "!", "?", and "&", respectively. Rendezvous events represent successful concurrent composition of corresponding send and receive events (*co-events*) having the same *names* (name[!e] = name[?e] = e). For example, the concurrent composition of send event !d and receive event ?d results in the rendezvous event &d.

Three operations in ACE describe composition of behaviors to form new behaviors. We introduce the first two next. The third, concurrent composition, is then shown to be a composite operator that can be defined in terms of the first two.

Sequential composition: if *e* is an event and E a behavior, then $e \bullet E$ means the occurrence of *e* followed by *E*.

Non-deterministic composition: if A and B are two behaviors, then A + B means the occurrence of either A or B non-deterministically. The sigma symbol is used as an n-ary generalization of + operator.

Two expressions E_1 and E_2 in ACE are said to be *equivalent*, $E_1 \equiv E_2$, if and only if, they represent the same event sequences and the choices at corresponding nodes in their Behaviors are the same. The meaning of the operations in ACE can be described using the following *axioms*:

Axioms:

A1. $A + B \equiv B + A$	*commutative property*
A2. $A + (B + C) \equiv (A + B) + C$	*associative property*
A3. $A + \$ \equiv A$	*identity property*
A4. $A + A \equiv A$	

Concurrent composition: if A and B are two behaviors and *scope*(A,B) is the set of *names* of events with which they communicate, then A | B is the new behavior of their concurrent execution. This can be simulated by having each pair of co-events produce a rendezvous event, and considering all possible shuffling of these events.

Let $A = \sum_{i=1}^{n} a_i \bullet A_i$ and $B = \sum_{j=1}^{m} b_j \bullet B_j$. Then the concurrent composition behavior of A and B can be defined by

$$A|B \;=\; \sum a_i \bullet (A_i | B) \quad + \quad \sum b_j \bullet (A | B_j) \quad + \quad \sum \&e \bullet (A_i | B_j)$$

$\forall\, a_i$, *names* $(a_i) \notin$ *scope* (A,B)

$\forall\, b_j$, *names* $(b_j) \notin$ *scope* (A,B)

$\forall\, a_i$ and b_j where:
name $(a_i) = (b_j) = e \in$ *scope*(A,B) ;
and a_l and b_j are *co-events*.

Concurrent composition is easily shown to be commutative and associative. Also, from A3 and the definition above, we can deduce that A|\$ = A.

We can now define the syntax of an expression in ACE. Let the set of events be denoted by ζ and the set of identifiers be denoted by Γ. Let the set of expressions in ACE be denoted by χ. Upper case letters will be used to range over χ.

Definition 1: An expression $E \in \chi$ in ACE is either: \$, $I \in \Gamma$, $e \bullet E$ $(e \in \zeta)$, E + E, or, E|E.

The behavior of each protocol entity can be specified by a *complete* set of equations each of the form (I = E, where $I \in \Gamma$ and $E \in \chi$) in which all behavior identifiers are specified. A protocol specification then consists of the specification of all entities in the protocol and the scope of communication between each pair of communicating entities. Consider a basic data transfer protocol whose configuration is shown in Fig. 1 with three communicating entities: sender S, full-duplex channel M, and a receiver R. The sender sends a message (!d) and waits to receive an acknowledgment (?d')[1] from the receiver before sending another message. An internal timeout event (&t) is used to counter situations where messages are lost in which case the message is resent. The receiver sends an acknowledgment (!a) when a message arrives (?d'). The channel delivers messages and acknowledgments or could lose messages (&l) assuming that acknowledgments cannot be lost in this example for simplicity reasons. The specification of the data transfer protocol described above is as follows:

PROTOCOL BasicDataTransfer *S, M, R*

scope(S,M) = {*d, a'*}
scope(M,R) = {*d', a*}
scope(S,R) = ∅

ENTITY S
S = !d•S1
S1 = ?a'•S + &t•S2
S2 = !d. •S1 + ?a'. •S

ENTITY M
M = ?d•M1
M1 = !d'•M2 + &l•M4
M2 = ?a•M3 + ?d•M1
M3 = !a'•M
M4 = ?d•M1

ENTITY R
R = ?d'•R1
R1 = !a•R

END_PROTOCOL

---

[1] We use d' (a') to denote the name of the corresponding event d (a) at the other end of the channel.

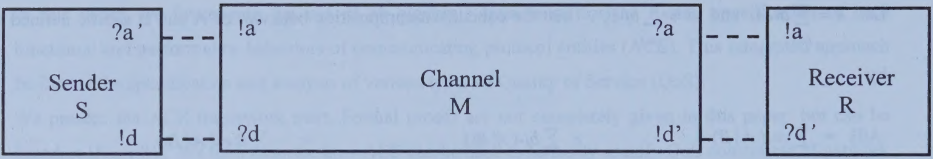


Figure 1: Basic Data Transfer Protocol

The concurrent composition of S and M produces a new composite entity SM as shown in Fig. 2. SM is then concurrently composed with R to produce the overall protocol concurrent behavior SMR as shown in Fig. 3.

SM = S | M
SM = &d•S1M1
S1M1 = !d'•S1M2 + &l•S1M4 + &t•S2M1
S1M2 = &t•S2M2 + ?a•S1M3
S1M4 = &t•S2M4
S2M1 = !d'•S2M2 + &l•S2M4
S2M2 = ?a•S2M3 + &d•S1M1
S1M3 = &a'•SM + &t• S2M3
S2M4 = &d•S1M1
S2M3 = &a'•SM

Figure 2: Concurrent Composition of S and M

SMR = SM | R
SMR = &d•S1M1R
S1M1R = &d'•S1M2R1 + &l•S1M4R + &t•S2M1R
S1M2R1 = &t•S2M2R1 + &a•S1M3R
S1M4R = &t•S2M4R
S2M1R = &d'•S2M2R1 + &l•S2M4R
S1M3R = &a'•SMR + &t•S2M3R
S2M2R1 = &a•S2M3R + &d•S1M1R1
S2M4R = &d•S1M1R
S2M3R = &a'•SMR
S1M1R1 = &l•S1M4R1 + &t•S2M1R1
S1M4R1 = &t•S1M1R1
S2M1R1 = &l•S2M4R1
S2M4R1 = &d•S1M1R1

Figure 2: Concurrent Composition of SM and R

Specification of Functional Constraints

We will define two functions, *Terminate* and *Precedence*, to map a given behavior to some sub-behavior. *Terminate* can be used to get a partial behavior terminating with some given event and identifier, or to get a terminating behavior from a cyclic one. A behavior is *terminating* if and only if one of its summands is of the form *a* • $, where *a* is any event, or of the form *a* • B where B is terminating; otherwise it is *cyclic.*

The concurrent behavior SMR is cyclic and it describes the behavior of the protocol through the message delivery phase and acknowledgement delivery phase. We will derive its sub-behavior for the course of one cycle starting from the initial state *SMR* and ending with the delivery of the message, i.e., event &d'. Let this terminating behavior be denoted by SMR_T.

Let Π represent the power set of a set that includes all pairs of events and identifiers (e, I), where $e \in \zeta$ and $I \in \Gamma$.

Definition 2: *Terminate* is a function $\chi \times \Pi \rightarrow \chi$ such that

$$Terminate[\sum_{i=1}^{n} a_i \bullet A_i, P \in \Pi] = \sum_{i \neq j} a_i \bullet Terminate[A_i, P] + \sum_{\forall (aj, Aj \in P)} a_j \bullet \$$$

SMR_T is formally specified by

$SMR_T = Terminate[SMR, \{(\&d', {*}M2R1)\}]$

where * stands for S, S1, or S2. This produces the following terminating behavior

$\mathrm{SMR}_T = \&\mathrm{d} \bullet \mathrm{S1M1R}_T$

$\mathrm{S1M1R}_T = \&\mathrm{d}' \bullet \$ + \&\mathrm{l} \bullet \mathrm{S1M4R}_T + \&\mathrm{t} \bullet \mathrm{S2M1R}_T$

$\mathrm{S1M4R}_T = \&\mathrm{t} \bullet \mathrm{S2M4R}_T$

$\mathrm{S2M1R}_T = \&\mathrm{d}' \bullet \$ + \&\mathrm{l} \bullet \mathrm{S2M4R}_T$

$\mathrm{S2M4R}_T = \&\mathrm{d} \bullet \mathrm{S1M1R}_T$

Precedence can be used to map a given behavior to a sub-behavior by pruning some of the summands. This is performed based on given pairs of events such that whenever the events of a given pair are contending, then the summand of the second event in the pair is eliminated. In such event pairs, the first event is said to *have precedence* over the second. We use *Precedence* to derive the sub-behavior SMR_p of SMR_T where there are no premature timeouts, i.e. timeout would not occur if the timeout event &t is contending with any other event in SMR_T.

Let Φ denote the power set of the set of all event pairs.

Definition 3: *Precedence* is a function: $\chi \times \Phi \rightarrow \chi$ such that

$$Precedence[A = \sum_{i=1}^{n} a_i \bullet A_i, S \in \Phi] = \sum_{i=1, i \neq j}^{n} a_i \bullet Precedence[A_i, S] \qquad \forall (a_k, a_j) \in S; a_k, a_j \in CH(A)$$

SMR_p is formally specified by

$SMR_p = Precedence[SMR_T, \{(\&l, \&t), (\&d, \&t), (\&d', \&t)\}]$

$SMR_p = \&\mathrm{d} \bullet \mathrm{S1M1R}_p$

$\mathrm{S1M1R}_p = \&\mathrm{d}' \bullet \$ + \&\mathrm{l} \bullet \mathrm{S1M4R}_p$

$\mathrm{S1M4R}_p = \&\mathrm{t} \bullet \mathrm{S2M4R}_p$

$\mathrm{S2M4R}_p = \&\mathrm{d} \bullet \mathrm{S1M1R}_p$

Note that given a protocol behavior, the functions above can be used to map it to sub-behaviors that are often much smaller in terms of the number of equation. A protocol designer, to save time and effort, may find it attractive in some instances to concentrate on these sub-behaviors instead of the complete protocol behavior. The defined sub-behaviors will be used in specifying performance features next and also in specifying QoS later.

Performance Specification

Consider the timing of a protocol behavior specified by a set of algebraic equations of the form $A = \sum a_i \bullet (A_i | B)$. Assume that all the expressions are *well-formed, i.e.* all events a_i, i=1 to n, in each expression are distinct. The *derivative* of such well formed expressions A relative to an event a, denoted by $\vartheta_a(A)$, is defined to be equal to A' if and only if there is a summand of A of the form $a \bullet A'$; otherwise, it is undefined.

Let us observe the timing behavior of A starting from some initial time t_0. At this time the choice set $CH(A) = [a_1,\dots,a_n]$ is said to be *enabled*, meaning that any event belonging to this choice set may occur at the next occurrence time t_1. If event a_i occurs at t_1, then the choice set $CH(\vartheta a_i(A))$ is enabled. This continues until termination of A when the choice is $CH(\$) = \varnothing$ is enabled and no events can possibly occur. Generally, if an expression E_{i-1} is observed at t_{i-1} ($E_0 = A$), then any event $a \in CH(E_{i-1})$ may occur at t_i. The choice set $CH(\vartheta_a(E_{i=1}))$ is then enabled at t_i. Only events belonging to a choice set enabled at t_{i-1} can occur at t_i. Also, only one choice set is enabled at any time instant. Note that each occurrence time t_i has a collection of expressions and choice sets associated with it for every event that may occur at t_{i-1}.

Such protocol timing behavior can be modeled as a *marked point entity* [Snyd 75] $[\{t_i, m_i\}, i \geq 0]$, where t_i is the *i*-th occurrence time and m_i is the *i*-th *mark* denoting the set of possible events that may occur at t_i. Each mark m_i is a collection of choice sets $CH(E_{i-1})$ for each expression E_{i-1} associated with t_{i-1} ($m_0 = \varnothing$).

Let us assume that the occurrence times of a protocol timing behavior are *continuous random variables.* Computing statistics of these occurrence times is very complex because they are dependent on the marks associated with them and the event that occurred at the previous occurrence time. The computations can be simplified by assuming that the occurrence time of an event, is *independent* of the other events in the same choice set. The occurrence time of an event is measured from the time its choice set is enabled until it occurs.

Suppose that expression C is observed at some time t_{i-1}. Let τ_a denote the occurrence time of event $a \in CH(C)$ measured from t_{i-1}, and $F_a(t)$ denote the probability distribution of τ_a. Then, the probability that such event a occurs at t_i, and the mean and variance of the time until it occurs are given by the following lemma.

Lemma 1

1. The probability that event $a \in CH(C)$ occurs at t_i conditioned on the fact that $CH(C)$ is enabled at t_{i-1}, is given by

$$\rho_C(a) = \int_0^{\infty} \prod_{a_i \in CH(C), a_i \neq a} [1 - F_a(t)] \, d F_{ai}(t) \qquad (1)$$

The time duration from t_{i-1} when $CH(C)$ is enabled until $a \in CH(C)$ occurs (at t_i) has a conditional mean, denoted by $\mu_C(a)$, and variance, denoted by $\sigma_C(a)$, which are given by

$$\mu_C(a) = \int_0^\infty \prod_{a_i \in CH(C), a_i \neq a} [1 - F_{ai}(t)]dt \qquad (2)$$

$$\sigma_C(a) = \int_0^\infty t \prod_{a_i \in CH(C), a_i \neq a} [1 - F_{ai}(t)]dt - \mu_C^2(a) \qquad (3)$$

Proof: An outline of the proof is as follows. The events belonging to a choice set $CH(C)$ enabled at t_{i-1} are actually *contending* with each other for occurrence at t_i. However, only one of them can occur since when one occurs, for example a, it then enables the choice set $CH(\partial_a(C))$. This implicitly means that the choice set $CH(C)$ is disabled since only one choice set can be enabled at any one time. As a consequence of the contention between the events for occurrence at t_i and their mutual exclusion, the conditional probability of a occurring is given by

$$\rho_C(a) = \Pr(a \text{ is } Min\{\tau_a, \forall a_j \in CH(C)\}) \qquad (4)$$

This leads to the first equation in the lemma. From the contention of events for an occurrence time and their mutual exclusion, the time duration from t_{i-1} to t_i is given by

$$t_{i-1} - t_i = Min\{\tau_{a_j}, \forall a_j \in CH(C) \qquad (5)$$

The mean and variance of this time lead to the second and third equations in the lemma, respectively.

Let C and A be terminating and well-formed expressions such that A represents a subset of the summands of C, to be denoted by *Sum*(C). A *probability attribute* $\mathbf{P}_C(A)$ denotes the conditional probability of A occurring relative to the sample space *Sum*(C), given that the choice set of C is enabled. Let the *time duration* of A relative to C be defined as the length of time starting when *CH*(C) is enabled until A terminates. A *mean-time attribute* $\mathbf{M}_C(A)$, and a *variance-time attribute* $\mathbf{V}_C(A)$ are the mean and variance of the duration time of A relative to C, respectively.

Note that each of these attributes is a function from terminating and well-formed expressions in ACE, denoted by $\chi_t \subset \chi$, to positive real numbers: $\chi_t \times \chi_t \rightarrow \Re^+$ (in the case of the probability attribute, the real numbers are bounded by 0 and 1). Theorems 1, 2, and 3 given next define mappings from operations in ACE to arithmetic operations on the probability, mean-time, and variance time attributes, respectively.

Theorem 1

P1. $\mathbf{P}_C(a \bullet A) = \rho_C(a) \cdot \mathbf{P}_{\partial_a(C)}(A)$

P2. $\mathbf{P}_C(\sum_{i=1}^{n} a_i \bullet A_i) = \sum_{i=1}^{n} \mathbf{P}_C(a_i \bullet A_i)$

Theorem 2

$$\text{M1.}\quad \mathbf{M}_C(a \bullet A) \;=\; \mu_C(a) \cdot \mathbf{M}_{\partial_a(C)}(A)$$

$$\text{M2.}\quad \mathbf{P}_C\Big(\sum_{i=1}^{n} a_i \bullet A_i\Big) \;=\; \frac{\sum_{i=1}^{n} \rho_C(a_i) \cdot \mathbf{M}_C(a_i \bullet A_i)}{\sum_{i=1}^{n} \rho_C(a_i)}$$

The only assumptions made so far regarding the timing behavior of protocols are that the occurrence times are continuous random variables and that the occurrence times of contending events are independent. No assumptions were made regarding their probability distributions. Hence, the above theorems for computing the probability and mean-time attributes of protocol timing model can be applied to any distribution of the events occurrence times for which a first moment exists. However, with such a general assumption, the rules for computing the variance-time attribute would be considerably complex; in particular, the computation of $\mathbf{V}_C(a \bullet A)$. Therefore, for the sake of simplifying these computations, the occurrence times of events will be assumed throughout the rest of the dissertation to be *exponentially* distributed. Such exponentially distributed occurrence times of event are independent on the time from which they are enabled. Consequently, a particular event would have just one value for the rate for its occurrence time.

Theorem 3

$$\text{V1.}\quad \mathbf{V}_C(a \bullet A) \;=\; \sigma_C(a) \cdot \mathbf{V}_{\partial_a(C)}(A)$$

$$\text{V2.}\quad \mathbf{V}_C\Big(\sum_{i=1}^{n} a_i \bullet A_i\Big) \;=\; \frac{\sum_{i=1}^{n} \rho_C(a_i) \cdot \big[\mathbf{V}_C(a_i \bullet A_i) + \mathbf{M}_C(a_i \bullet A_i)^2\big]}{\sum_{i=1}^{n} \rho_C(a_i)} - \mathbf{M}_C^2\Big(\sum_{i=1}^{n} a_i \bullet A_i\Big)$$

Let λ_a denote the exponential rate of event a. In terms of the exponential rates of the occurrence time of events $\rho_C(a)$, $\mu_C(a)$, and $\sigma_C(a)$ of lemma 1 are reduced to

$$\rho_C(a) = \frac{\lambda_a}{\sum_{a_i \in CH(C)} \lambda_{a_i}} \qquad \text{if and only if } a \in CH(C) \tag{6}$$

$$\mu_C(a) = \frac{1}{\sum_{a_i \in CH(C)} \lambda_{a_i}} \qquad \text{if and only if } a \in CH(C) \tag{7}$$

$$\sigma_C(a) = \frac{1}{\Big(\sum_{a_i \in CH(C)} \lambda_{a_i}\Big)^2} \qquad \text{if and only if } a \in CH(C) \tag{8}$$

By inspection, it is clear that the rules for computing the timing attributes of given well-formed and terminating expressions respect axioms A1-A3 of ACE. That is, given A ≡ B, then $\mathbf{P}_C(A) = \mathbf{P}_C(B)$, $\mathbf{M}_C(A) = \mathbf{M}_C(B)$, and $\mathbf{V}_C(A) = \mathbf{V}_C(B)$. This is not true though for Axiom A4 (A+A ≡ A) since the timing attributes are not defined for A+A, which is not well-formed.

Corollary 1

Let $A = a \bullet A + B$, then

$$\mathbf{M}_A(A) = \frac{\rho_1(a)}{[1-\rho_1(a)]} \cdot \mu_A(A) + \mathbf{M}_A(B) \tag{9}$$

$$\mathbf{V}_A(A) = \frac{\rho_1(a)}{[1-\rho_1(a)]} \cdot \sigma_A(a) \quad \frac{\rho_1(a)}{[1-\rho_1(a)]^2} \cdot \mu_A^2(a) + \mathbf{V}_A(B) \tag{10}$$

Given a set of algebraic equations describing a protocol behavior, the rules in theorems 1, 2, and 3 can be used to map them into a set of linear equations. In these linear equations, the variables are the attributes and the coefficients are arithmetic expressions in the event rates.

Specification QoS Requirements

Any QoS requirement that can be specified in terms of the attributes of the timing model of protocols, can be computed using the mapping rules of theorems 1, 2, and 3, and lemma 1. Only an algebraic specification of the protocol and the exponential rates of events involved in it are used in these computations.

For the data transfer protocol, we will formally specify four QoS constraints. First, to limit the probability of message loss to some given value ε, the following constraint can be stated

QoS1. $\lambda_{\&l} < \varepsilon$

Second, to keep the mean transfer time below a given threshold t, we can specify the following constraint

QoS2. $M_{SMR_T}(SMR_T) < t$

Third, to limit the jitter in the transfer time to some small value j, we can specify the following constraint

QoS3. $V_{SMR_T}(SMR_T) < j$

Fourth, to minimize the possibility of network congestion due to premature timeouts, we can limit the probability of sub-behavior SMR_p to a small value p with the following constraint

QoS4. $P_{SMR_T}(SMR_P) < p$

Conclusion

An algebraic framework, ACE, for the integrated specification of functional and performance requirements of network protocols was presented and used to formally specify QoS requirements. The framework lends itself to automation, which would facilitate early investigation of various protocol designs and parameters to meet given QoS. We assumed exponential rates for all protocol events and demonstrated the feasibility of the framework to specify QoS requirements that can be stated in terms of mean, variance, and probability attributes of protocol behaviors and sub-behaviors.

References

[Camp94] A. Campell, G. Coulson and D. Hutchison. A Quality of Service Architecture. ACM Computer Communication Review, April 1994.

[Holz91] G. Holzmann. Design and Validation of Computer Protocols. Prentice-Hall, 1991.

[ISO88] ISO, ISO 8807, Information Processing Systems, Open System Interconnection, LOTOS-A Formal Technique Based on the Temporal Ordering of Observational Behavior, 1988.

[ISO89] ISO, ISO 8807, Information Processing Systems, Open System Interconnection Estelle: A Formal Description Technique Based on an Extended State Transition Model, 1989.

[ITUT93] ITU-T. Recommendation Z.100: Specification and Description Language (SDL). Geneva, Switzerland, 1993.

[Noun86] N. Nounou. A Methodology for Specification-Based Performance Analysis of Communication Protocols. Ph.D. thesis, computer science department, Columbia University, 1986.

[Snyd75] D. Snyder. Random Point Processes. Wiley-Interscience, 1975.

Analysis of an FEC/NAK Approach for Scalable Reliable Multicast

Oliver Z. Zeng, Chilukuri K. Mohan, and Carlos R. P. Hartmann
Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NY 13244-4100
{zzeng, mohan, hartmann}@top.cis.syr.edu

Abstract

This paper presents quantitative analyses and numerical results that support the case that FEC should be incorporated into IP multicast to achieve reliability and scalability. We analyze and compare two receiver-initiated reliable multicast models: an integrated FEC/NAK model and a NAK based ARQ model. We show that the use of parity packets in FEC/NAK will greatly reduce packet loss probability, and as the number of receivers increases, the effect of reducing packet loss probability will result in dramatic gains over NAK based ARQ in the expected number of transmissions, transmission efficiency, bandwidth loss, and protocol scalability.

1 Introduction

IP multicast provides efficient routing and delivery of IP packets to multiple receivers. But it does not provide reliability, since it uses the same best-effort delivery semantics as other IP datagram delivery. Multicast datagrams can be lost, delayed, or duplicated. Reliability is a critical concern for distributing information such as stock quotes, medical images and software packages. Since the inception of IP multicast, a number of protocols have been proposed to provide reliability on top of IP multicast. Most of these reliable multicast protocols were designed for dozens or hundreds of receivers, using Automatic Repeat reQuest (ARQ) based schemes to achieve reliability. But they do not easily scale up to a larger number of receivers because of the state explosion and the message implosion problems [8]. This paper demonstrates that this problem can be alleviated by incorporating Forward Error Correction (FEC) into reliable multicast protocols.

In other work on the use of FEC in multicasting, Metzner [4] studied the use of FEC/ARQ for reliable multi-point transmission and suggested the use of Reed-Solomon codes for parity computation. McAuley [3] pointed out that multicasting may be the most beneficial application in using FEC and proposed a burst erasure correcting Reed-Solomon code as a complement to ARQ for broadband communication. Huitema [2] performed quantitative analyses to demonstrate why FEC gain outweighs its pain when multicasting to large groups. Nonnenmacher *et al.* [5, 6] have done performance modelling of reliable multicast transmission with respect to different topologies.

This paper presents further analytical and numerical results that make a compelling case for using FEC in reliable multicast. Both [2] and [5] have shown that by incorporating FEC, the expected number of transmissions decreases when the number of receivers increases. [2] also depicted how transmission efficiency varies with packet loss rate, while [5] analyzed how the use of FEC affects the achievable end-host throughput. Besides the expected number of transmissions, we compare transmission efficiency, bandwidth loss, and scalability improvement for a transmission group of k packets (vs. a given packet in [2] and [5]) on two receiver-initiated reliable multicast models: a negative acknowledgement (NAK) based ARQ model and an integrated FEC and NAK model (FEC/NAK). Our analysis shows that as the number of receivers increases, FEC/NAK performs much better in all four respects.

The rest of the paper is organized as follows. Section 2 lays the foundation of our analysis by describing a simplified multicast tree model and two reliable multicast schemes above it: a NAK based ARQ and an integrated FEC/NAK. The expected number of transmissions for any given packet, transmission efficiency, and bandwidth loss for both schemes are measured and compared in Section 3, 4, and 5, respectively. Section 6 quantifies the improvement in scalability for reliable multicast using FEC. Section 7 presents the concluding remarks.

2 A Simplified Analysis Frame

Analysis will be conducted within the framework of a complete binary tree as the topology for multicast connections [6], and two reliable multicast schemes built upon the tree: an FEC/NAK protocol and its counterpart, a NAK based ARQ. We assume that packet loss for each receiver is independent and homogeneous [1, 10].

We compare a NAK based ARQ (that exemplifies receiver-initiated reliable multicast protocols, e.g., [7] and [9]) with an integrated FEC and NAK protocol. For the simplicity of protocol description and analysis without losing generality, we assume there are k source packets need to be multicast to all the R receivers. Using the term in [5], these k source packets are referred to as a *transmission group*. We seek to define schemes that exhibit the following common behaviors:

- All packets are multicast by the sender.
- The sender adds a sequence number in each packet transmitted. When it has no packets to transmit, it sends a heartbeat packet containing the last sequence number. By looking at the sequence number, the receiver can detect if it has lost any packets.
- NAKs are transmitted to the sender over a unicast channel, and no NAK suppression is applied.
- NAKs are never lost, since control packets are small.

In NAK based ARQ scheme, whenever a receiver loses a given packet, it asks for retransmission of the packet by sending a NAK. In FEC/NAK, at the sender, h parity packets are computed using a Reed-Solomon erasure code such as the one described by McAuley [3], so that reception of any unique k of the encoded n $(n = k + h)$ packets can be used to reconstruct the original k packets when packet loss happens. Then the original k packets as well as the redundant h packets are sent. If at least k packets are received, no retransmission is needed. Lost original packets are reconstructed and delivered. A receiver will only ask for retransmission of the lost original packets by NAK when less than k packets are received. Clearly packet loss rate is reduced in FEC/NAK. We will quantify this effect and show how much improvement it will make on multicasting performance.

Suppose the independent packet loss rate is p. For FEC with $k+h$ encoded packets, the event that a packet is lost will happen only if it is not received and fewer than k out of the remaining $k+h-1$ packets are received. So the packet loss rate in FEC/NAK equivalent to that in NAK based ARQ is:

$$p_h = p\left(1 - \sum_{i=k}^{k+h-1} \binom{k+h-1}{i} (1-p)^i p^{k+h-1-i}\right)$$

For example, if packet loss rate $p = 1\%$, adding 1, 2, or 3 parity packets to 8 original packets, will reduce p to much smaller packet loss rate: $p_1 \approx 7.73 \times 10^{-4}$, $p_2 \approx 3.44 \times 10^{-5}$, or $p_3 \approx 1.14 \times 10^{-6}$.

For simplicity, we will compare FEC/NAK schemes of multicasting 8+1, 8+2, and 8+3 packets with NAK based ARQ scheme of multicasting only the original 8 packets from Section 3 to 6.

3 Expected Number of Transmissions

The expected number of transmissions to deliver a packet to all receivers is an important measure in reliable multicast. It captures the global packet loss behavior in the tree, the cost of processing plus the delay of a reliable multicast delivery. It also lays the foundation for calculating transmission efficiency, bandwidth loss, and amount of feedback.

Let N represent the number of transmissions required to transmit a given packet to all R receivers. The probability that all R receivers require exactly i transmissions of a given packet is: $Pr(N = i) = (1 - p^i)^R - (1 - p^{(i-1)})^R$. For NAK based ARQ, the expected number of transmissions of a packet is:

$$E(N) = \sum_{i=1}^{\infty} i Pr(N = i)$$

For FEC/NAK, for every k source packets to be multicast, a total of $k + h$ packets are sent. Taking into account transmissions due to redundancy, the expected number of transmissions for a given packet is:

$$E_h(N) = \frac{k+h}{k} \sum_{i=1}^{\infty} i\left((1 - p_h^i)^R - (1 - p_h^{(i-1)})^R\right)$$

It can be shown that $E(N) = \sum_{j=1}^{R} \binom{R}{j} (-1)^{(j+1)} \frac{1}{1-p^j}$. To obtain numerical values for $E(N)$ and $E_h(N)$, we truncated the series $\sum_{i=1}^{\infty} i Pr(N = i)$ when $Pr(N = i) \leq 10^{-7}$.

Figure 1 shows that NAK based ARQ requires more transmissions than all three FEC/NAK schemes when $R \geq 9$. Unlike [2, 5], our comparisons are more meaningful since we use the same number of original packets to compare NAK based ARQ with FEC/NAK.

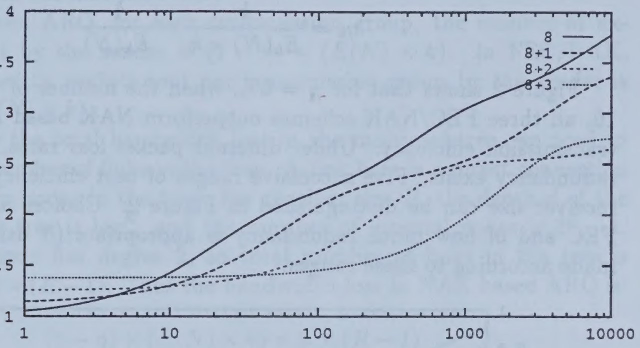


Figure 1: Expected number of transmissions vs. R ($p = 5\%$)

Shortened end-to-end delay can also be expected due to the decrease in the expected number of transmissions.

4 Transmission Efficiency

Transmission efficiency is defined as the "goodput" ratio, i.e., the number of packets received and processed by each receiver divided by the total number of packets transmitted by the sender. It is an important measurement that indicates the percentage of bandwidth being used for useful transmissions. For point to point communication, FEC/NAK is always less efficient than NAK based ARQ, transmission wise, since in FEC redundant packets are always sent with original packets while in NAK based ARQ they are transmitted only after errors have occurred. However, the scenario changes in multicast transmission.

Transmission efficiency can be quantified based on the expected number of transmissions. For a transmission group of k packets, the efficiency for

NAK based ARQ is:

$$\eta = \frac{k}{E(N) \times k} = \frac{1}{E(N)}$$

For FEC/NAK, in general with $k + h$ scheme:

$$\eta_h = \frac{k}{E_h(N) \times k} = \frac{1}{E_h(N)}$$

Figure 2 shows that for $p = 5\%$, when the number of receivers exceeds 10, all three FEC/NAK schemes outperform NAK based ARQ in terms of transmission efficiency. Under different packet loss rates, no best level of redundancy exists. Five successive ranges of best efficiency with respect to receiver size can be distinguished in Figure 2. Choices of whether to use FEC and of how much redundancy is appropriate (if using FEC) can be made according to these ranges.

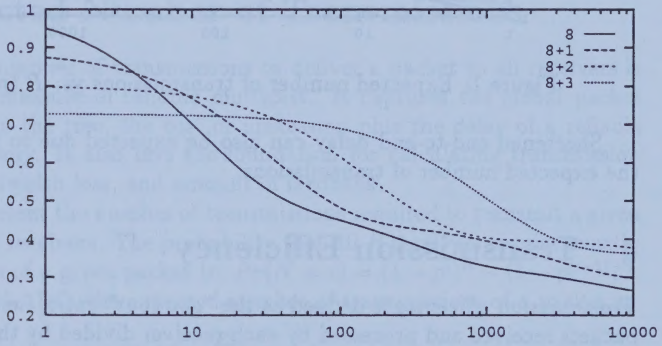


Figure 2: Transmission efficiency vs. R ($p = 5\%$)

5 Bandwidth Loss

For NAK based ARQ, bandwidth loss only comes from packet retransmission. In FEC/NAK, redundancy adds the other source of bandwidth loss. Since packet retransmission in FEC/NAK occurs much less often due to the effect of reduced packet loss rate, bandwidth saving in FEC brought by redundancy will in fact significantly exceed the bandwidth loss because of the redundancy.

In this section, we estimate bandwidth loss as the total number of extra packets sent per transmission group over the links of the complete binary tree. We first quantify the number of extra packets sent per transmission group for both NAK based ARQ and FEC/NAK, then we calculate the total bandwidth loss in the multicast tree.

In NAK based ARQ, for each transmission group, the number of extra packets sent by the sender is $(1 - \eta) \times (E(N) \times k)$. In FEC/NAK, the number of extra packets sent per transmission group by the sender is $(1 - \eta_h) \times (E_h(N) \times k)$.

To calculate the total bandwidth loss in the multicast tree, we need to determine the number of links in the tree. A link here is a direct connection between any two nodes in the tree. For receiver size R, the depth d of the complete binary tree is $\log_2^R$, and the number of internal nodes is $2^d - 1$. Each internal node has degree 2, so total number of links in the tree is $2 \times (2^d - 1) = 2 \times (R - 1)$. Thus the bandwidth loss in NAK based ARQ is:

$$(1 - \eta) \times (E(N) \times k) \times 2 \times (R - 1)$$

The bandwidth loss in FEC/NAK is:

$$(1 - \eta_h) \times (E_h(N) \times k) \times 2 \times (R - 1)$$

In Figure 3, for multicasting 8 packets to a group of 10^4 receivers under packet loss rate of 5%, adding 3 parity packets will in fact result in a bandwidth saving of about 220000 packets.

6 Scalability

In a NAK based ARQ protocol, a packet will be retransmitted if it is lost by just one receiver. The probability of this event, $1 - (1 - p)^R$, grows with the size R of the receiver group, which leads to the reception of more and more NAKs and the retransmission of an increasingly larger fraction of the original packets. Furthermore, since the sender needs to know the sequence numbers of lost packets to decide which packets to retransmit, this will lead to the so-called "message implosion" [8] and cause serious scalability problems.

These problems can be greatly alleviated in FEC/NAK. Incorporating FEC into the protocol has two effects on the reduction of feedback. First, due to the significantly reduced packet loss rate at the receivers by applying

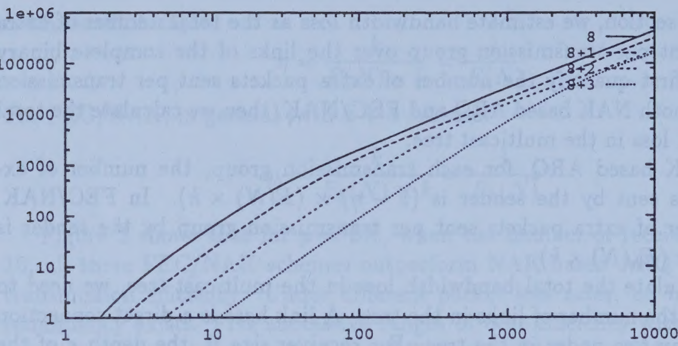


Figure 3: Bandwidth loss (number of packets) vs. R ($p = 5\%$)

FEC, the amount of feedback is very much reduced. Second, because of the property of erasure codes, the sender needs only to know how many packets are lost by any receiver but not the sequence number of each of the lost packets. Feedback is further reduced from the per-packet scenario in NAK based ARQ to the per-transmission group scenario in FEC/NAK.

We now quantify the above analysis for both models. Without NAK suppression, the expected number of NAKs returned by a given receiver for a packet is $1/(1-p) - 1 = p/(1-p)$. In NAK based ARQ, for a transmission group of k packets, the amount of feedback from R receivers is:

$$R \times k \times \frac{p}{1-p}$$

In FEC/NAK, p is decreased to p_h and per-packet feedback is reduced to per-transmission group feedback, so the amount of feedback from R receivers is:

$$R \times \frac{p_h}{1-p_h}$$

Figure 4 shows how feedback is dramatically reduced from NAK based ARQ to FEC/NAK for the three simple FEC cases when the size R of the receiver group goes up to 10^4.

On the one hand, FEC adds complexity to the sender, but on the other hand, it greatly reduces the complexity of feedback (NAK) processing on the sender. Reducing the amount of feedback is an extremely useful feature for multicast protocols since it gives protocols the potential to scale well to a large number of receivers.

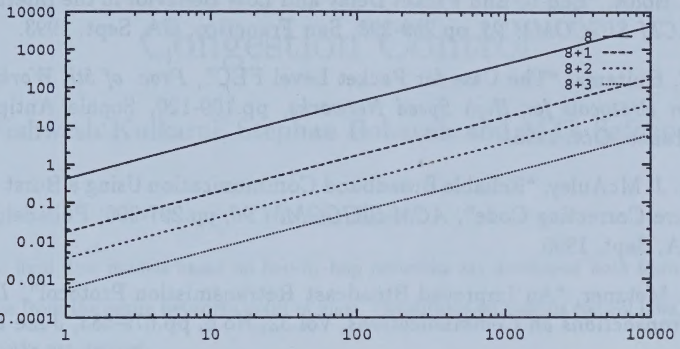


Figure 4: Amount of feedback vs. R ($p = 5\%$)

7 Conclusion

We have presented quantitative analyses and numerical results that support the case of incorporating FEC into IP multicast to achieve reliability and scalability. Our analyses examine two receiver-initiated reliable multicast models: an integrated FEC/NAK model and a NAK based ARQ model. We show that the use of parity packets in FEC/NAK will greatly reduce packet loss probability. As the number of receivers increases, the effect of reducing packet loss probability will result in dramatic gains over NAK based ARQ in the following respects: (1)Decrease in expected number of transmissions for a given packet; (2)Increase in transmission efficiency; (3)Decrease in bandwidth loss; and (4)Improvement in protocol scalability.

We conclude that FEC should be incorporated into NAK based protocols to achieve scalable reliable multicast transmissions.

Acknowledgement

The authors thank Professor Kishan Mehrotra for help in solving $E(N)$ in Section 3.

References

[1] J. Bolot, "End-to-End Packet Delay and Loss Behavior in the Internet", *ACM SIGCOMM'93*, pp.289-298, San Francisco, CA, Sept. 1993.

[2] C. Huitema, "The Case for Packet Level FEC", *Proc. of 5th Workshop on Protocols for High Speed Networks*, pp.109-120, Sophia Antipolis, France, Oct. 1996.

[3] A. J. McAuley, "Reliable Broadband Communication Using a Burst Erasure Correcting Code", *ACM SIGCOMM'90*, pp.297-306, Philadelphia, PA, Sept. 1990.

[4] J. Metzner, "An Improved Broadcast Retransmission Protocol", *IEEE Transactions on Communications*, Vol.32, No.6, pp.679-683, June 1984.

[5] J. Nonnenmacher, E. W. Biersack, and D. Towsley, "Parity-based Loss Recovery for Reliable Multicast Transmission", *ACM SIGCOMM'97*, Cannes, France, September 1997.

[6] J. Nonnenmacher and E. W. Biersack, "Performance Modelling of Reliable Multicast Transmission", *IEEE INFOCOM'97*, Kobe, Japan, April 1997.

[7] M. Handley and J. Crowcroft, "Network Text Editor (NTE): A Scalable Shared Text Editor for the Mbone", *ACM SIGCOMM'97*, Canne, France, August 1997.

[8] D. Towsley, J. Kurose, and S. Pingali, "A Comparison of Sender-initiated and Receiver-initiated Reliable Multicast Protocols", *IEEE Journal on Selected Areas in Communications, Special Issue on Network Support for Multipoint Communications*, Vol.15, No.3, April 1997.

[9] X. R. Xu, A. C. Myers, H. Zhang, and R. Yavatkar, "Resilient Multicast Support for Continuous-Media Applications", *NOSSDAV'97*, 1997.

[10] M. Yajnik, J. Kurose, and D. Towsley, "Packet Loss Correlation in the MBone Multicast Network", *IEEE GLOBCOM'96*, London, Nov. 1996.

Stability Issues in Hop-by-Hop Rate Based Congestion Control

Vishwesh Kulkarni, Stephan Bohacek and M.G. Safonov

I. Abstract

Continuous in time fluid flow models based on hop-by-hop networks are developed both from a single connection's oint of view as well as from the entire networks point of view. Conditions on various control laws that would guarantee tability of such networks are derived.

II. Introduction

The objective of flow control in data networks is to provide a reasonable trade-off between network throughput and uality of service - usually a function of the number of packet losses and end-to-end delay/jitters. The state-of-the-art eployed feedback control mechanisms in the Internet are *end-to-end* flow control methods since the control is enforced nly at the source end-host and the intermediate routers on the way to the destination end-host do not take part in ow control (apart from scheduling), e.g. the feedback algorithms used in TCP-suites [1], [2],[3]. There are also *explicit* eedback methods in which a router sets a *congestion bit* in the packet header if its queue length exceeds a certain limit s in DEC-bit [4]. In addition, *hop-by-hop* (HBH) rate based congestion control [5] has also been of interest as a research roblem. In HBH, a feedback signal signifying congestion at a node is relayed to the nodes which are sending data to it. n response to the congestion signal received, a node updates its sending accordingly. It has been speculated that HBH s inherently unstable though simulation results for some particular HBH schemes argue against it [5]. The objective of his paper is to study stability of continuous in time fluid flow models based on such networks.

he paper is organized as follows. In Section III, network model from the point of view of a particular connection is escribed. In Section IV, stability of the model is examined. In Section V, a network model without any concern for particular connection is obtained and its stability is examined.

III. Problem Formulation

Let us consider a particular unicast connection $C_{s,d}$ which has source s and destination d. Let there be a fixed path rom s to d during the lifetime of $C_{s,d}$. Let there be N routers, denoted by r_1 through r_N, on path from s to d. Let the ervice capacity of r_i at time t be given by $\tilde{\mu}_i(t)$, which w.l.o.g. can be taken to be a time-invariant $\tilde{\mu}_i$. Let the service apacity of r_i at time t for the rest of the connection be $\tilde{\mu}_i^d(t)$ $(\leq \tilde{\mu}_i)$. Then the effective service capacity of r_i for $C_{s,d}$

This research is supported by AFOSR F49620-98-1-0026.

The authors are with Department of Electrical Engineering, University of Southern California, Los Angeles CA90089-2563.
mail: vishwesh@usc.edu, bohacek@usc.edu, safonov@usc.edu

at time t is given by

$$\mu_i(t) = \tilde{\mu}_i(t) - \tilde{\mu}_i^d(t). \tag{1}$$

We call r_i *upstream* of r_j if its index in the ordered set $\{s, r_1, \dots, r_N, d\}$ of the nodes of $C_{s,d}$ is smaller than r_j. Let $\tilde{q}_i(t)$ and $\bar{q}_i$ denote, respectively, the actual and desired queue sizes at r_i at time t. Let the sending rate at r_i for $C_{s,d}$, denoted by $x_i(t)$, be given by $\tilde{x}_i(t)$. Let the desired (pre-set) rate be $\bar{x}_i(t) \doteq \bar{x}(t) = \bar{x}\ \forall i, \forall t$. The rationale for pre-set rates is as follows. A connection i is set up only when the service rate for the connection is guaranteed to lie within $[\bar{x} - \delta, \bar{x} + \delta]$ for some $\bar{x}, \delta$. Thus, in the face of new connections coming up or old connections getting terminated, the service rate stays within bounds. Define

$$q_i(t) \doteq \tilde{q}_i(t) - \bar{q}_i(t) \qquad \forall i = 1, 2, \dots, N \tag{2}$$

$$x_i(t) \doteq \tilde{x}_i(t) - \bar{x}_i(t) \qquad \forall n = 1, 2, \dots, N. \tag{3}$$

Let us assume dynamics to be linear about $\{x_i, q_i | i = 1, 2, \dots, N\}$. Let r_i update $x_i(t)$ according to the following law.

$$\dot{x}_i(t) = \lambda x_i(t) + k\, q_{i+1}(t) \qquad \forall n = 1, 2, \dots, N. \tag{4}$$

The queue dynamics are given by

$$\dot{q}_i(t) = \tilde{x}_{i-1}(t) - \tilde{x}_i(t) + u_i(t) \qquad \forall i = 1, 2, \dots, N-1 \tag{5}$$

where $u_i(t)$ denotes the difference between incoming traffic and outgoing traffic at r_i due to the rest of the connections. Let $x_0(t)$ define the rate of source at time t. Define $\tilde{x}(t) \doteq [x_0(t)\ x_1(t)\ \dots\ x_N(t)\ q_1(t)\ \dots\ q_{N-1}(t)]^T$. For notational simplicity, let us refer to $\tilde{x}$ by x. Define $u(t) \doteq [u_1(t)\ u_2(t)\ \dots\ u_{N-1}(t)]^T$. Then (4)-(5) can be rewritten as

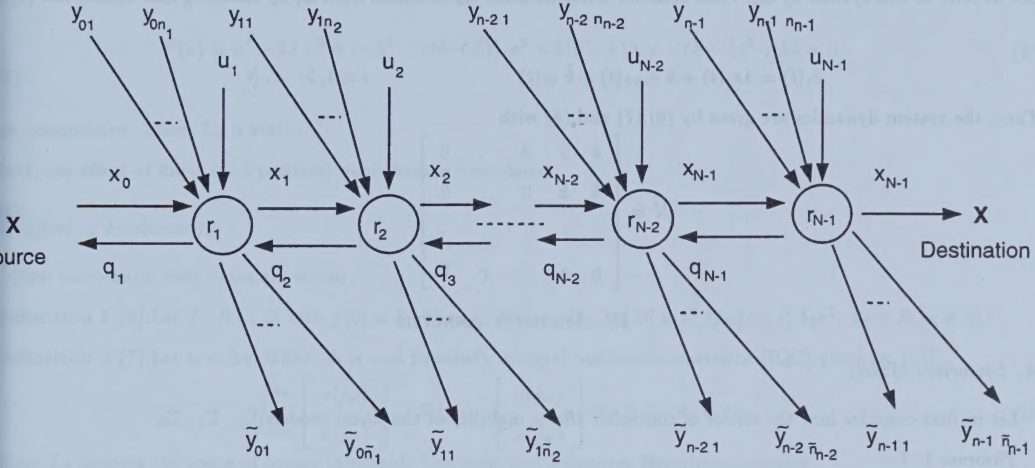
$$\dot{x}(t) = Ax(t) + Bu(t) \tag{6}$$

where

$$A = \begin{bmatrix} \Lambda & K \\ W & 0 \end{bmatrix}, \quad B = [\underbrace{0, 0, \dots, 0}_{N}, 1, 1, \dots, 1] \tag{7}$$

$$\Lambda = \lambda \cdot I, \quad K = k \cdot I, \quad W = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

and I denotes $N \times N$ identity matrix and 0 denotes $N \times N$ matrix with all entries equal to zero. Let us refer to the system presented by (6) by Σ_1. Σ_1 is a simplified model of the system in which the effect of the dynamics of connections hich share links with $C_{s,d}$ is taken to be that of a disturbance. If a state is included to account for each of such flows, odel Σ_2 is obtained (see Fig. 1). Let $P_i \doteq \{y_{i,1}, y_{i,2}, \dots, y_{i,n_i}\}$ denote the set of incoming rates at r_{i+1} which share



y01
y0n1
y11
y1n2
yn-2 1
yn-2 nn-2
yn-1
yn-1 nn-1
u1
u2
uN-2
uN-1
---
x0
x1
x2
xN-2
xN-1
xN-1
X
ource
r1
r2
rN-2
rN-1
X
Destination
q1
q2
q3
qN-2
qN-1
ỹ01
ỹ0ñ1
ỹ11
ỹ1ñ2
ỹn-2 1
ỹn-2 ñn-2
ỹn-1 1
ỹn-1 ñn-1

g. 1. The figure shows routers r_i exchange feedback information q_i to update data rates x_i for a particular connection $C_{s,d}$. The figure represents Σ_2 with y_{ij}, $\tilde{y}_{ij}$ being the incoming and outgoing flows due to other connections and u_i account for the effect of connections coming up or getting terminated. By setting y_{ij}, $\tilde{y}_{ij}$ to zero, Σ_1 is obtained.

e link r_i-r_{i+1} with $C_{s,d}$. Let $\tilde{y}_{i,1}, \tilde{y}_{i,2}, \dots, \tilde{y}_{i,\tilde{n}_i}$ ($\tilde{n}_i \le n_i$) denote the rates from P_i which do not share link $r_{i+1} - r_{i+2}$ ith $C_{s,d}$. Define $Y_i \doteq [y_{i,1}\ y_{i,2}\ \dots\ y_{i,n_i}]^T$, $Z_i \doteq [\tilde{y}_{i,1}\ \tilde{y}_{i,2}\ \dots\ y_{i,\tilde{n}_i}]^T$. Write $n_x \doteq N + \sum_{i=0}^{N-2} n_i + \tilde{n}_i$. Define

$$x \doteq [x_0\ Y_0\ Z_0\ x_1\ Y_1\ Z_1\ \dots\ x_{N-1}\ q_1\ q_2\ \dots\ q_{N-1}]^T \tag{9}$$

$$\Lambda \doteq diag(\lambda,\ \lambda,\ \dots,\ \lambda) \in R^{n_x \times n_x} \tag{10}$$

$$K_i \doteq [\underbrace{k\ k\ \dots\ k}_{n_i}\ 0\ 0 \dots\ 0]^T \qquad i = 0, 1, \dots, N-1 \tag{11}$$

$$\phi_i \doteq [0\ 0\ \dots\ 0] \in R^{\sum_{j=0}^{i-1} n_i + \tilde{n}_i} \tag{12}$$

$$W_i \doteq [\phi_i, \underbrace{1,\ 1,\ \dots,\ 1}_{1+n_i}, \underbrace{-1,\ -1,\ \dots,\ -1}_{1+\tilde{n}_i},\ 0,\ \dots, 0] \tag{13}$$

$$W = [W_1; W_2; \dots; W_N], \quad K \doteq [K_0\ K_1\ \dots\ K_{N-1}]; \tag{14}$$

hen, system dynamics are given by

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{15}$$

here

$$A = \begin{bmatrix} \Lambda & K \\ W & 0 \end{bmatrix} \tag{16}$$

Let us refer to this system by Σ_2. Next consider a modification Σ_3 obtained from Σ_1 by changing rate update law (4) to

$$\dot{x}_i(t) = \lambda x_i(t) + k\ q_{i+1}(t) + \bar{k}\ q_i(t) \qquad\qquad i = 1, 2, \ldots, N. \tag{17}$$

Then, the system dynamics are given by (6),(7) and (8) with

$$K \doteq \begin{bmatrix} k & 0 & 0 & \ldots & 0 \\ \bar{k} & k & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & \bar{k} \end{bmatrix}, \tag{18}$$

IV. STABILITY ANALYSIS

A. Linearized Models

Let us first consider how the choice of controller affects stability of the linear models Σ_1, Σ_2, Σ_3.

Theorem 1: Let

$$l \leq 0, \quad k \leq 0, \tag{19}$$

$$|l| \geq |\sqrt{l^2 + 4k}| \qquad\qquad \text{if } l^2 + 4k \geq 0 \tag{20}$$

Then, Σ_1 is stable. If either $l > 0$ or $k > 0$, the system is unstable.

Proof. Let ν be an eigenvalue of A and let v be its corresponding right eigenvector. Then

$$A\ v = \nu\ v \Longleftrightarrow \begin{cases} \lambda\ v_i + k\ v_{i+N} = \nu\ v_i & \text{i=1,2,\ldots,N} \\ v_i - v_{i+1} = \nu\ v_{i+N} & \text{i=1,2,\ldots,N-1.} \end{cases} \tag{21}$$

Solving (21), it can be verified that the eigenvalues of A are $\nu = \lambda$ and $\nu = \frac{\lambda \pm \sqrt{\lambda^2 + 4k}}{2}$. Σ_1 is unstable if an eigenvalue of A lies in open right half s-plane. Σ_1 is stable if all of the eigenvalues of A lie in open left half s-plane. From these definitions, (19) and (20) follow. QED.

Theorem 2: Let (19) and (20) hold. Then, Σ_2 is stable. If either $l > 0$ or $k > 0$, the system is unstable.

Proof. It can be shown on the lines of the proof for Theorem 1 that the eigenvalues of matrix A of Σ_2 are $\nu = \lambda, \frac{\lambda \pm \sqrt{\lambda^2 + 4n_i k}}{2}$ where n_i is as defined before. Note that since $k \leq 0$

$$\sqrt{\lambda^2 + 4n_i k} < \sqrt{\lambda^2 + 4k} \tag{22}$$

$$\Longleftrightarrow \frac{\lambda \pm \sqrt{\lambda^2 + 4k}}{2} < 0 \Longrightarrow \frac{\lambda \pm \sqrt{\lambda^2 + 4n_i k}}{2} < 0. \tag{23}$$

Hence Σ_2 is stable if (19)-(20) are satisfied and is unstable otherwise. QED.

Theorem1 and Theorem2 together show that from stability point of view, it is sufficient to check the stability of Σ_1 rather than consider the more involved model Σ_2. On similar lines the following theorem can be proved.

Theorem 3: Let $\lambda, k, \bar{k}$ be such that the roots of

$$P(x) = x^4 - 2\lambda\, x^3 + (-\lambda^2 - 2(k - \bar{k}))\, x^2 + 2(k - \bar{k})\lambda\, x + (k - \bar{k})^2 + k\bar{k} = 0 \tag{24}$$

are nonpositive. Then, Σ_3 is stable.

Next, the effect of finite (and positive) queuesizes is considered.

B. Effect of Nonlinearity

First some basic definitions are stated.

Definition 1 [6]: Let $f : R \mapsto R$ with $f(0) = 0$. Then, $f \in \text{sector}[k_1, k_2]$ iff $k_1 e^2 \le e f(e) \le k_2 e^2,\ \forall e \in R\ (e \neq 0)$.

Definition 2 [7]: Let $u = \Delta v$. Then, Δ is said to satisfy integral quadratic constraint (IQC) given by Π if

$$\int_{-\infty}^{\infty} \begin{bmatrix} v(j\omega) \\ u(j\omega) \end{bmatrix}^{*} \Pi(j\omega) \begin{bmatrix} v(j\omega) \\ u(j\omega) \end{bmatrix} d\omega \ge 0 \quad \forall v \in L_2 \tag{25}$$

where L_2 denotes the space of square integrable functions and $*$ denotes Hermitian operator.

In reality, $0 \le \tilde{q}(t) \le q^* \ \forall t$. Because of this saturation nonlinearity, the pure integral action in (5) is replaced by $\dot{q}(t) = f(\zeta)$ where $\zeta(t) = \tilde{x}_{i-1}(t) - \tilde{x}_i(t) + u_i(t)$ and

$$f(\zeta) = \begin{cases} 1 & \text{if } |\tilde{q}(t)| \le q^* \\ & \text{or if } \tilde{q}(t) = q^* \text{ and } sgn(u) = -1 \\ & \text{or if } \tilde{q}(t) = -q^* \text{ and } sgn(u) = 1 \\ 0 & \text{else.} \end{cases} \tag{26}$$

Note that $f \in \text{sector}[0, 1]$. Let hop-to-hop delay be $\tau\ \ (0 \le \tau \le \tau^*)$. This modeled as $(\Delta u)(t) = u(t - \tau^*)$. For simplicity let us consider Σ_4 with source and destination separated by 3 routers. Tapping signals u_i and v_i as shown in fig 2, the plant P and uncertainty Δ decomposition is given as $v = P\ u,\ \ u = \Delta\ v$ where

$$P = \begin{bmatrix} 0 & 0 & -K(s) & 0 & K(s)/s & 0 \\ 0 & 0 & 0 & -K(s) & \bar{K}(s)/s & K(s)/s \\ 0 & 0 & 0 & 0 & 1/s & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/s \\ -1 & 0 & -K(s) & K(s) & (K(s) - \bar{K}(s))/s & -K(s)/s \\ 0 & -1 & 0 & -K(s) & \bar{K}(s)/s & (K(s) - \bar{K}(s))/s \end{bmatrix} \tag{27}$$

$$\Delta = diag(1 - e^{-\tau s},\ 1 - e^{-\tau s},\ 1 - e^{-\tau s},\ 1 - e^{-\tau s},\ f(\cdot),\ f(\cdot)). \tag{28}$$

For N hop case, the form of Δ is obvious and P can be seen to be a $3N \times 3N$ matrix with $N-1 \times N-1$ component matrices (see Fig 2) bearing relationship with 2×2 counterparts in (2) as

$$\begin{bmatrix} K(s)/s & 0 \\ \bar{K}(s)/s & K(s)/s \end{bmatrix} \mapsto \begin{bmatrix} K(s)/s & 0 & 0 & 0 & \dots & 0 \\ \bar{K}(s)/s & K(s)/s & 0 & 0 & \dots & 0 \\ 0 & \bar{K}(s)/s & K(s)/s & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \bar{K}(s)/s & K(s)/s \end{bmatrix} \quad \text{etc.}$$

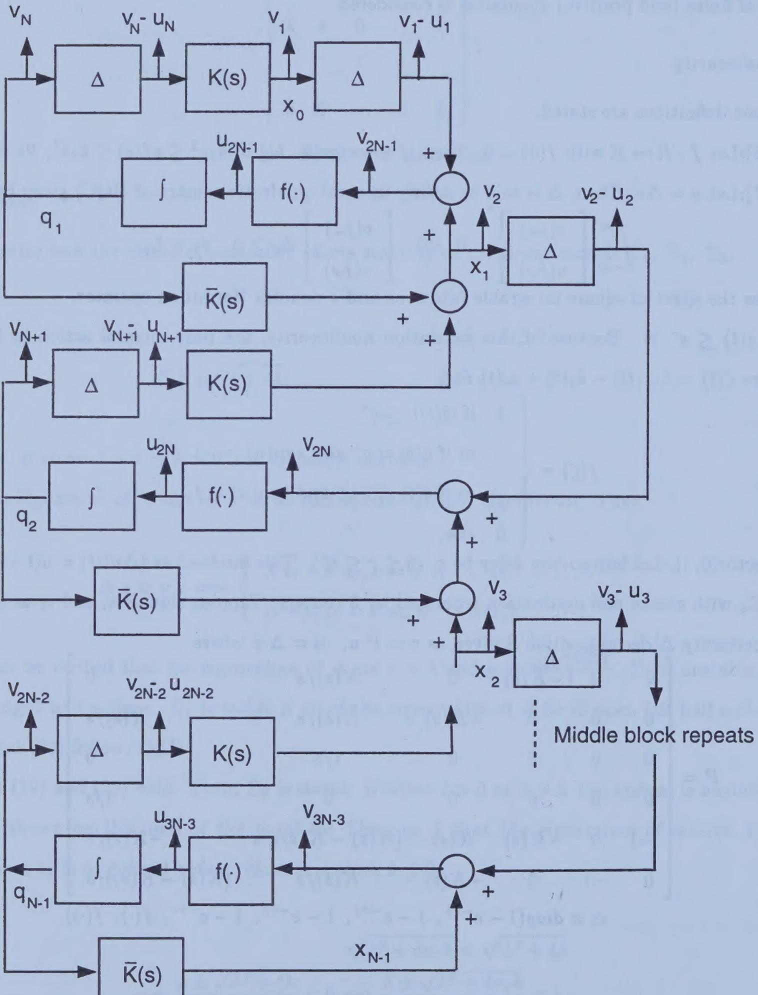

V_N
$V_N - U_N$
V_1
$V_1 - U_1$
Δ
K(s)
Δ
X_0
u_{2N-1}
V_{2N-1}
+
∫
f(·)
V_2
$V_2 - U_2$
q_1
+
Δ
X_1
$\bar{K}$(s)
+
+
V_{N+1}
$V_{N+1} - U_{N+1}$
Δ
K(s)
u_{2N}
V_{2N}
∫
f(·)
q_2
+
+
$\bar{K}$(s)
V_3
$V_3 - U_3$
+
+
Δ
X_2
V_{2N-2}
$V_{2N-2} - U_{2N-2}$
Δ
K(s)
Middle block repeats
u_{3N-3}
V_{3N-3}
∫
f(·)
q_{N-1}
+
+
$\bar{K}$(s)
X_{N-1}

Fig. 2. The figure shows tapping of signals used to decompose the system into uncertainty Δ and plant P seen by Δ.

Using IQCs for the uncertainties (see [8] for a list of IQCs), sufficiency conditions for the stability of the system can be derived. If τ is assumed to be negligible, as would be the case with high-speed data links, $\Delta = diag(f(\cdot),\ f(\cdot))$ and satisfies IQC given by

$$\Pi(j\omega) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -2 & 0 \\ 0 & 1 & 0 & -2 \end{bmatrix}. \tag{29}$$

Theorem 4: Σ_4 is stable if $K(s), \bar{K}(s)$ are chosen such that

$$\Big(\Big(\frac{K(s)-\bar{K}(s)}{s}\Big)^* + \Big(\frac{K(s)-\bar{K}(s)}{s} - 2\Big)\Big)^2 - \Big|\Big(\frac{\bar{K}(s)-K(s)}{s}\Big)\Big|^2 < 0. \tag{30}$$

Proof: It can be seen that the plant P in this case is

$$P = \begin{bmatrix} (K(s)-\bar{K}(s))/s & -K(s)/s \\ \bar{K}(s)/s & (K(s)-\bar{K}(s))/s \end{bmatrix}. \tag{31}$$

For stability, it is sufficient if [7]

$$\begin{bmatrix} P \\ I \end{bmatrix}^* \Pi \begin{bmatrix} P \\ I \end{bmatrix} < 0 \quad \text{on } s = jR. \tag{32}$$

Expanding and using Schur's compliment [9], the result follows. Q.E.D.

V. Model of the Entire Network

Now we extend the continuous time, linear system model and controller to the entire network.

A. *System model*

Define the matrix $[\alpha_{i,j}]$ by

$$\alpha_{i,j} = \begin{cases} 1 & \text{if router } i \text{ sends data to router } j \\ 0 & \text{otherwise} \end{cases}$$

et q_i denote the queue size at r_i, $\bar{q}_i$ denote the short term average queue size r_i, $x_{i,j}$ denote the rate at which data s being sent from r_i to r_j, $\bar{x}_{i,j}$ denote the short term average rate at which data is sent from r_i to r_j, $v_{i,j}$ denote he deviation from current distribution, d_in_i denote the rate at which data enters the network (from the un-modeled etwork) at r_i, d_out_i denote the rate at which data leaves the network (to the un-modeled network) at r_i and let $_i \doteq d\_in_i - d\_out_i$. The queue dynamics are

$$\dot{q}_i = -\sum_j \alpha_{i,j} x_{i,j} + \sum_j \alpha_{j,i} x_{j,i} + u_i. \tag{33}$$

ince $x_{i,j}$ is not defined for each i, j we defined $\alpha_{i,j} x_{i,j}$ to be zero if $\alpha_{i,j} = 0$. We impose the following rate update law

$$\dot{x}_{i,j} = -\lambda\,(x_{i,j} - \bar{x}_{i,j}) - k_1\,(q_j - \bar{q}_j) + k_2\,(q_i - \bar{q}_i) + v_{i,j} \tag{34}$$

here, for simplicity, $k_1, k_2, \lambda > 0$ are constant. For the moment, assume that $v_{i,j} = 0$ As will be shown, the $\lambda\,(x_{i,j} - \bar{x}_{i,j})$ term forces the system to be stable. From (34), it is clear that a steady state is achieved when $x_{i,j} = \bar{x}_{i,j}$ nd $q_j = \bar{q}_j$ for all i, j. That is, the rate $x_{i,j}$ is the same as $\bar{x}_{i,j}$. Since, it is not known what the average rate at any uter should be, we cannot predefine $\bar{x}_{i,j}$. Thus, we simply let $\bar{x}_{i,j}$ to be a low pass filtered version of $x_{i,j}$. Likewise, is a low pass version of q_j. In particular,

$$\begin{aligned} \dot{\bar{x}}_{i,j} &= -\delta\bar{x}_{i,j} + \delta x_{i,j} \\ \dot{\bar{q}}_{i,j} &= -\gamma\bar{q}_{i,j} + \gamma q_{i,j} \end{aligned} \tag{35}$$

where γ and δ are small. The exogenous inputs u_i are used to account for the interaction of the modeled network with un-modeled network - either routers or end-hosts. By definition, if $u_i > 0$, then more data is entering the network at this router than is leaving. At the steady state, the same amount of data enters the network as it leaves. Thus $\sum_i u_i = 0$. Now suppose that a new connection begins with data entering the network at r_j and leaving the network at r_k. Furthermore, before this new connection, suppose that no data was entering or leaving the network through these routers, i.e. $u_j = u_k = 0$. As the connection begins, u_j becomes positive. But, since the data has not yet reached r_k, $u_k = 0$, thus $\sum_i u_i > 0$. In time, the data reaches router k and $u_k < 0$ since by assumption the rest of the connections are operating at steady state, so that $\sum_i u_i = 0$. After the connection terminates, $u_j = 0$. However, until the last packet of data has reached r_k, we have $u_k < 0$ and $\sum_i u_i < 0$. Finally, when the last packet reaches r_k and $u_k = 0$ and $\sum_i u_i = 0$. Thus

$$\int_{-\infty}^{\infty} \sum_i u_i = 0 \tag{36}$$

but $\int_{t_1}^{t_2} \sum_i u_i$ might not be zero for a choice of t_1, t_2 $(t_1 < t_2)$. We assume that the connections are set so that the data enters and leaves at the same rate. The exact value of this rate is not be discussed.

The input u does not account for all the variability that can occur. In particular, u can remain constant and the connections may change. The inputs $v_{i,j}$ account for these other changes. Define $xq_{i,j}$ to be the rate at which data is entering r_i with destination r_j. Define some nonlinear system f such that when

$$v_{i,j} = f\left(x_{i,j}, \sum_k \alpha_{i,k} x_{i,k}, xq_{i,j}, \sum_k \alpha_{i,k} xq_{i,k}, q_i\right)$$

the short term average of the fraction of data that leaves r_i for r_j is the same as the short term average of the fraction of data entering the r_i that is to be sent to r_j. One possibility is

$$v_{i,j} = \frac{dxq_{i,j}}{dt} - \left(-\lambda \tilde{x}_{i,j} - k_1\left(q_j - \bar{q}_j\right) + k_2\left(q_i - \bar{q}_i\right)\right).$$

For our purposes, the exact form of $v_{i,j}$ is not important. However, it is important to note that

$$\sum_j v_{i,j} = 0 \qquad \text{for all } i. \tag{37}$$

That is, if the fraction of data destined for one router increases, the fraction must decrease for some other routers. Furthermore, $v_{i,j}$ are not used to model an increase in the flow of data leaving the router. The total flow of data out of a router can only change by some queues filling up or emptying.

B. Stability of Network

It will be proven that with the rate update law defined by equation (34), the entire network is stable. That is, the linear system mapping inputs u and v to the queuelengths $q_i - \bar{q}_i$ and data rates $x_{i,j} - \bar{x}_{i,j}$ are asymptotically stable. After some manipulation, the transfer function from u_i and $v_{i,j}$ to $\tilde{q}_i$ is found to be

$$\frac{s+\delta}{s\left(s+\lambda+\delta\right)}\left(\frac{s\left(s+\gamma\right)\left(s+\lambda+\delta\right)}{s+\delta}I - \left(k_1 P_1 + k_2 P_2\right)\right)\tilde{q} = u + \left(\frac{\delta+s}{s\left(s+\lambda+\delta\right)}\right)v$$

where P_1 and P_2 are some matrices, $\tilde{q}_i \doteq q_i - \bar{q}_i$, $\tilde{q} \doteq [\tilde{q}_1 \cdots \tilde{q}_n]^T$, $u \doteq [u_1 \cdots u_n]^T$, $v \doteq [\sum_j \alpha_{j,1} v_{j,1} \cdots \sum_j \alpha_{j,n} v_{j,n}]^T$. The poles of the linear system which has input u and ouput $\tilde{q}$ correspond to the values of s such that

$$\left(\frac{s\left(s+\gamma\right)\left(s+\lambda+\delta\right)}{s+\delta}I - \left(k_1 P_1 + k_2 P_2\right)\right) \tag{38}$$

is not invertible. Clearly this matrix is not invertible if and only if $\frac{s(s+\gamma)(s+\lambda+\delta)}{(s+\delta)}$ is an eigenvalue of $(k_1 P_1 + k_2 P_2)$. The next lemma describes the eigenvalues of $(k_1 P_1 + k_2 P_2)$.

Lemma 1: The eigenvalues of $(k_1P_1 + k_2P_2)$ are real non-positive. There is one eigenvalue at zero. The eigenvector ssociated with the zero eigenvalue is the constant vector, i.e. $\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T$.

ince $(k_1P_1 + k_2P_2)$ has an eigenvalue at zero, the matrix (38) is not invertible for $s = 0$. Thus the linear system as a pole at zero. However, since u is multiplied by s, this pole is not excited by u. The other poles are the values f $s \neq 0$ where s is such that $\frac{s(s+\gamma)(s+\lambda+\delta)}{s+\delta}$ is an eigenvalue of $(k_1P_1 + k_2P_2)$. By Lemma 1, s must be such that $\frac{s+\gamma)(s+\lambda+\delta)}{s+\delta} = -r$, for some $r \geq 0$. For $r = 0$, $s = 0$ corresponds to the pole at zero. For $r > 0$,

$$s\,(s+\gamma)\,(s+\lambda+\delta) + r\,(s+\delta) \;=\; 0 \tag{39}$$

pplying the Routh stability criterion, we find that since all the constants are positive, the solutions to (39) has negative al parts if $(\lambda+\gamma+\delta)\,(\gamma\,(\lambda+\delta)+r) - r\delta > 0$, which is true for all $r > 0$. Therefore, we conclude that all poles occur r values of s such that $Re(s) < 0$. Hence, the linear system from inputs u to $\tilde{q}$ is asymptotically stable.

ow we examine the mapping from v to q, that is

$$\left(\frac{(s+\gamma)\,(s+\lambda+\delta)\,s}{s+\delta} I - (k_1P_1 + k_2P_2)\right)\tilde{q} = \begin{bmatrix} \sum_j \alpha_{j,1}v_{j,1} \\ \vdots \\ \sum_j \alpha_{j,1}v_{j,n} \end{bmatrix}.$$

gain, there is a pole at zero that is not excited by the input v. This can be explained as follows: Setting $s = 0$, we amine the mapping

$$\left(-\,(k_1P_1 + k_2P_2)\right)\tilde{Q}\,(0) = V\,(0) \tag{40}$$

ere $\tilde{Q}\,(0)$ and $V\,(0)$ are the Laplace transform of $\tilde{q}$ and v evaluated at 0. That is,

$$V\,(0) = \begin{bmatrix} \int_{-\infty}^{\infty}\sum_j \alpha_{j,1}v_{j,1} & \cdots & \int_{-\infty}^{\infty}\sum_j \alpha_{j,n}v_{j,n} \end{bmatrix}$$

d $\tilde{Q}\,(0) = \begin{bmatrix} \int_{-\infty}^{\infty}\tilde{q}_1 & \cdots & \int_{-\infty}^{\infty}\tilde{q}_n \end{bmatrix}$. From (40)

$$\tilde{Q}\,(0) = (k_1P_1 + k_2P_2)^{-1}\,V\,(0) \tag{41}$$

ere $\tilde{Q}\,(0)$ is defined only if $V\,(0)$ is orthogonal to the direction that $(k_1P_1 + k_2P_2)^{-1}$ is infinite, that is, if $V\,(0)$ is hogonal to the eigenspace of zero eigenvalue of $(k_1P_1 + k_2P_2)$. By Lemma 1, this eigenspace is spanned by the vector $\cdots\ 1\Big]^T$. And by equation (37), $\sum_i\sum_j \alpha_{j,i}v_{j,i} = 0$. Thus $V\,(0)$ is orthogonal to $\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T$; the mapping) is well defined and $\tilde{Q}\,(0)$ is finite whenever is. Hence, the pole at zero is not excited by the input v.

he same way the transfer function from u to $\tilde{q}$ is stable, the other poles of the transfer function from v to $\tilde{q}$ are ble. Therefore, the linear system from u and v to $\tilde{q}$ is asymptotically stable. Since $\lambda > 0$, equation (34) implies, the ppings from u and v to $\tilde{x}$ are asymptotically stable. Of course, the mapping to q, $\bar{q}$, x and $\bar{x}$ are only marginally ble. For if u or v are nonzero, these values 'charge up' to some non zero value. This is to be expected, if $x_{i,j}$ is zero n no data flows. Furthermore, without direct control over u, we cannot make the queues go to zero.

VI. CONCLUSION

everal models of continuous in time fluid flow networks addressing the issues of hop-by-hop networks have been ied under different control strategies. The results characterize control laws purely from stability stand point.

REFERENCES

[1] Van Jacobson, "Congestion avoidance and control," *Proceedings of the ACM SIGCOMM '88*,pp. 314-329, Aug 1988.

[2] B.S.Brakmo and L.L. Peterson, "TCP Vegas: end to end congestion avoidance on a global Internet," *IEEE Journal on Selecte Areas in Communication,* vol. 13, No. 8, pp. 1465-1480,

[3] K. Fall and S. Floyd, "Simulation-based comparisons of Tahoe, Reno and Sack TCP", *Computer Communication Review,* vol 26, no. 3, pp. 5-21, July 1996.

[4] K. K. Ramakrishnan and R. Jain,"A binary feedback scheme for congestion avoidance in computer networks," *ACM Trans actions on Computer Systems*,vol. 8, no. 2, pp. 158-181, May 1990.

[5] P. P. Mishra, H. R. Kanakia and S. K. Tripathi, "On hop-by-hop rate based congestion control," *IEEE/ACM Transaction on Networking*,pp. 224-240, April 1996.

[6] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties,* Academic Press, New York 1975.

[7] A. Rantzer and A. Megretski, "System analysis via integral quadratic constraints," *Proceedings of the 33rd Conference o Decision and Control,* pp. 3062-3067, 1994.

[8] A. Megretski and A. Rantzer, "System analysis via integral quadratic constraints," *IEEE Transactions on Automatic Contro* vol. 42, no. 6, pp. 819-830, 1997.

[9] S. Boyd, L. Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in system and Control Theory,* SIAM, Philadel phia, 1994.

Topology Aggregation for Networks with Hierarchical Structure: A Practical Approach

Yuval Shavitt
Network & Service Management Research Dept.
Bell-Labs, Lucent Technologies
101 Crawfords Corner Rd., room 4G-627, Holmdel, NJ 07733-3030
shavitt@lucent.com

Abstract

This paper addresses the problem of aggregating the topology of a sub-network in a compact way. The problem arises from networks that have a hierarchical structure, where each sub-network must advertise the cost of routing between each pair of its border nodes. The straight-forward solution of advertising the exact cost for each pair has a quadratic cost which is not practical.

We seek an aggregation that minimizes the distortion, defined as the ratio between the advertised cost of traversing the network between two border node and the actual cost. We compare the average and maximum distortion of several heuristics, based on spanning trees, with a theoretically bounded aggregation scheme. We show that in practice, a combination of a minimum spanning tree with random spanning trees achieves better aggregation than a theoretically bounded aggregation scheme.

1 Introduction

Future global networks will consist of hierarchies of subnetworks, or domains. Such domains might, for example, correspond to those portions of the network that reside in a particular geographic region, or they may reflect the network management structure within organizations. Reasons of both scalability and security will mandate that domains not reveal the details of their internal structure to outside nodes [1, 6]. Instead, domains will disclose only a summary, or aggregated view, of the costs and availabilities associated with traversing them. Even if one considers scalability issues alone, the need for aggregation is obvious, since the computation and communication complexity of both source and hop-by-hop routing protocols grows at least linearly in the number of links in the network representation.

At one extreme, we might require that an aggregation scheme must not misrepresent *any* information about a domain's internal structure. In this case, the aggregated representations generated would have to include the costs of all possible transits through the domain. Such a representation may be viewed as a complete weighted graph whose vertices are the border nodes of the domain and whose edges represent the costs of the corresponding transits. Unfortunately, for additive parameters, such as delay, the size of such a representation grows quadratically in the number of border nodes, so this scheme, while absolutely faithful to the true structure of the domain, is not scalable enough to be useful in practice. For path parameters that are calculated by using the maximum or minimum functions on the path's link values, such as maximum available bandwidth, a linear representation by a tree is sufficient to represent the full details of the domain [7]. In this work, we concentrate on additive parameters.

If, however, we relax the requirement that the aggregated representation must be absolutely truthful, then many alternative schemes become feasible. The performance benefits of having a higher-fidelity aggregation scheme need to be balanced against the real-world burdens imposed by the use of larger representations. The burdens imposed by having larger, more accurate representations include greater space requirements for the topology databases within each domain, increased background traffic between domains due to topology updates, and longer computation times for determining the least-cost routes.

The benefit of larger, more accurate representations is principally better route selection. In PNNI [1], for example, a connection setup slated to travel through some domain (based on a source route calculation using *advertised* information) might discover upon arriving in the domain that the intended transit is actually not feasible. When this type of a blockage occurs (due to a discrepancy between truth and advertising) the connection setup packet must retrace its steps from the point of blockage back to its original entry point into the domain. The act of retracing one step is called a *crankback*. Clearly, lowering the frequency of crankbacks is a significant benefit of having higher fidelity aggregation schemes, since it results in faster setup times for connections, and less load on the network control.

Methods for compact, i.e., linear in the number of border nodes, graph representation that can be used to aggregate topology information were studied in the past years [9, 8, 2, 5]. These methods have a a logarithmic bound on the worst-case distortion of costs between any pair of nodes. In other words, for any pair of nodes the ratio between the advertised and the true cost of the shortest path that connect them is bounded by a $\log b$, where b is the number of border nodes. Awerbuch and Shavitt [4] presented an aggregation scheme based on Bartal trees [5] and showed that it achieves the promised logarithmic bound. Awerbuch et al. [3] conducted a simulation study of PNNI routing where they found that using a Minimum Spanning Tree (MST) aggregation does not degrade performance. Lee [7] suggested to use a maximum delay spanning tree as an upper bound estimate for the delay, which is an additive metric. The bounds calculated from his tree are not tight when the difference between the maximum and the minimum traversing delay is high, a situation occurring, e.g., when the original topology is sparse.

In this work, we investigate the aggregation performance of some heuristics based on MST and Random Spanning Tree (RST), and compare them with the t-tree aggregation suggested in [4], and with the maximum spanning tree suggested by Lee [7]. We show that MST is a good aggregation scheme, which corresponds to the simulations conducted in [3], and show that it can be enhanced by combining the MST with RSTs. We also show that a combination of three RSTs is very bad, which again corresponds to the results obtained in [3].

The rest of the paper is organized as follows. In the next section, we describe the PNNI standard that is used as our network model. In section 3, we describe previous relevant work. The suggested aggregation heuristics and their performance are described in the next section. Finally, we conclude by suggesting how to use our results in practice.

2 The PNNI Model

The PNNI standard is part of the ATM protocol suite that defines the network-to-network interface. It defines, among all, how the network is organized in a hierarchical manner into peer groups, how aggregated topology information is exchanged among peer groups, and how hierarchical source routing is performed. We selected this standard for our network model as it is a mature standard with a growing deployment base. However, our results hold in general for any topology aggregation, as our reliance on PNNI is very limited (see below).

In the PNNI standard [1, section 3.3.8], a network is aggregated ("complex node") by a star. A virtual node ("nucleus") is the interior reference point. For each state parameter, a single default value is given to represent the value of this parameter for the connection between a border node and the nucleus ("spoke") in each direction.

To cope with asymmetry, the standard allows "exceptions" of two types: a different value for a spoke of a certain border node, or a value for a "bypass" between two border nodes. In guidelines that are supplied with the standard [1, Appendix C] it is recommended that the number of exceptions used to configure a complex node will be kept smaller than $3b$, where b is the number of border nodes. There is no suggestion on how to select exceptions, or how to compute the default spoke value.

In our heuristics, we use only the default spoke value and up to $3b$ bypass exceptions. The default spoke value is the only part of our heuristic that might be considered, as specific to PNNI. A bypass exceptions is, in our view, a general way to represent any graph and we believe will be part of any future aggregation standard.

The performance measure used in this study is the *distortion*. The distortion between two border nodes is defined as the ratio between the cost of a parameter as calculated from the advertised topology and this cost in the real topology. As opposed to the usual practice in theoretical work [9, 8], we do not solely compare the worst case distortion, but examine also the average and the variance. It has been shown in the past [3] that worst case analysis cannot predict the efficiency of aggregation algorithms for routing in hierarchical networks.

3 Previous Work

Lee [7] suggests to use a maximum or minimum spanning tree to accurately represent path metrics that are calculated by taking the maximum or the minimum of the path's link values. E.g., he suggests to use maximum spanning tree to represent the maximum available bandwidth between any two nodes. For additive parameters, such as delay, he suggests to use a maximum spanning tree. The delay between any two points, u and v, that are not directly connected by the tree is estimated by the minimum delay of the links in the cycle generated when the link $(u,\ v)$ is added to the tree. If the network is highly symmetric his methods does not improve significantly the estimate given by the default spoke value. When the network is highly asymmetric the estimates are very bad. In particular, differences among the best routes (i.e., the shortest) tend to vanish, making the estimation not very useful. Our heuristic, attempts at reliably estimating the links with the low cost, while compromising accuracy for links with higher cost.

The calculation of the estimates using Lee's tree is problematic since it requires the receiving node to use a minimum function on the path between two nodes and not the standard addition. The use of the standard addition results in extremly bad estimates.

Thus Lee's suggestion cannot be incorporated into PNNI without some provisioning in the standard.

Bartal [5] suggests a tree representation for graphs with an average distortion of $O(\log n)$, where n is the number of nodes in the graph. The tree is not built from the edges and the nodes of the original graph. Instead, as many as $O(n)$ virtual nodes are required for this construction. If a Bartal tree is to be used to advertise a "complex node" the added virtual nodes should be advertised together with the link cost, which increases the representation cost of the "complex node".

Awerbuch and Shavitt [4] present a framework for aggregating a "complex node" topology based on Bartal trees. Their aggregation has a probabilistic bound on the worst case distortion in the advertised parameters which is $\log b$. They embed a variant of a Bartal tree in the network to keep the representation size of a single tree to be not higher than $b-1$. They suggest two structures based on the above embedded trees. A *t-tree*, which is the union of t random Bartal trees, and a slightly more complex construction called quality partition forest (QPF). They tested the t-tree construction and showed that the logarithmic bound is achieved in practice [4].

4 Heuristic Aggregation Algorithms

The construction of a Bartal tree (and thus of the t-tree aggregation) is complex and hard to code. This work suggests some simpler aggregation schemes that are based on spanning trees. Theoretically, an aggregation based on a tree have a maximal distortion which is linear in the graph size. For example, if a homogeneous ring (with the cost of all the links equal to 1) of size n is to be aggregated using a single tree the cost of the link which is removed is increased by a factor of $n-1$. However, the experiments we conducted and report here, show that, in practice, tree based aggregations have good distortion, better than the bounded aggregation of [4]. The good performance together with the ease of their implementation make them an attractive candidate as a heuristic method for practitioners.

We tested the following methods:

MST — the minimum spanning tree, comprised of $b-1$ edges.

3RST — three random spanning trees. This structure can have up to $3b-3$ edges. In practice, since the random trees overlap we used about $2.5b$ edges.

MST + 2RST — a MST plus two random spanning trees. The maximal number of edges here is $3b-3$, the actual number of edges was similar to the 3RST case.

***t*-tree** — several Bartal trees as described in [4]. In most cases, the number of edges used was just over $2b$.

Lee — a maximum spanning tree [7], where the distance evaluation is done by taking the minimum of the links connecting two nodes (see Section 3).

To compare between the aggregation schemes, we produce random graphs according to Waxman's method [10]. In the graph creation process, links are added until all nodes reach a minimum node degree of two. Links that increase node degree over 6 are rejected. This way the resulted graphs have both the clustering effect of close nodes and a relative high diameter. We used the graphs that contain 100 nodes. The number of border nodes

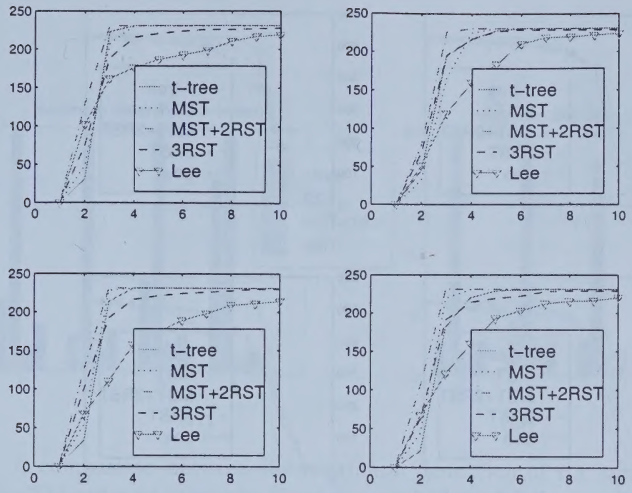


Figure 1: A comparison between the histogram of the aggregation schemes for the 22 border node graph.

was set to 22 and 33. For each graph we conducted several experiments, in which links were assigned different random values.

Figures 1 and 2 depict the cumulative histograms for the different aggregation schemes for four sets of experiments for each graph. The graphs show that MST has similar quality to the *t*-tree while using less than half the number of exceptions. MST+2RST has much tighter aggregation representation (with slightly more exceptions). In fact, in all but one experiment, this aggregation scheme had only a negligible number of node-pairs that were distorted by more than a factor of 2. 3RST performs similar to the *t*-tree, in most cases, but since its aggregation distortion is higher (and sometime by a large margin) than MST+2RST and since the calculation of a MST is not harder than the calculation of a RST, there is no reason to prefer this scheme. The histogram for the aggregation suggested by Lee [7] is, always, inferior to MST+2RST, and, in most cases, even inferior to MST.

Figures 3, 4, and 5 compare the maximum distortion, average distortion, and the variance of the distortion among the aggregation schemes. The average distortion of all the aggregation schemes is in the same range, but MST+2RST has the lowest average. The differences between MST+2RST and the rest of the aggregation schemes are between 20% and 100%, and up to 400% difference to Lee's method. However the maximum distortion graphs show that 3RST has a 1–2 orders of magnitude higher maximum distortion than the rest of the schemes. Lee's method produced even higher maximum distortion and thus was not plotted. In some of the cases, *t*-tree has a maximum distortion that is several times higher than that of MST. This result in a higher variance in the distortion for 3RST and *t*-tree. The addition of two RSTs to an MST practically reduces the variance to negligible values.

It is interesting to compare our results with the ones obtained by Awerbuch et al. [3]. In their simulation study, they compared the overall performance of the PNNI routing algorithm when several aggregation schemes are employed. The aggregation schemes studied in [3] include MST and RST, which are compared to the case where no

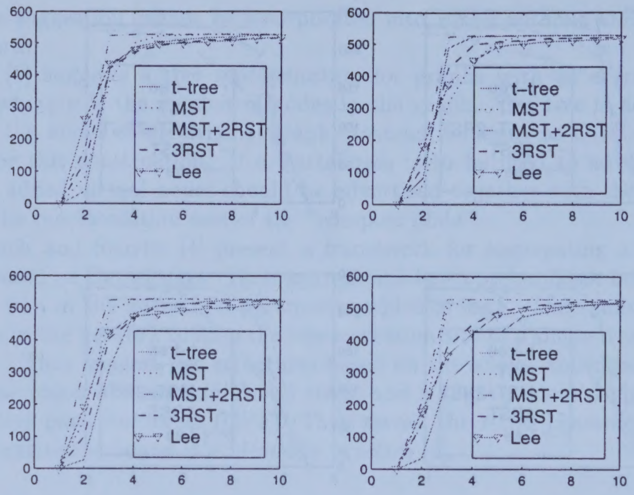
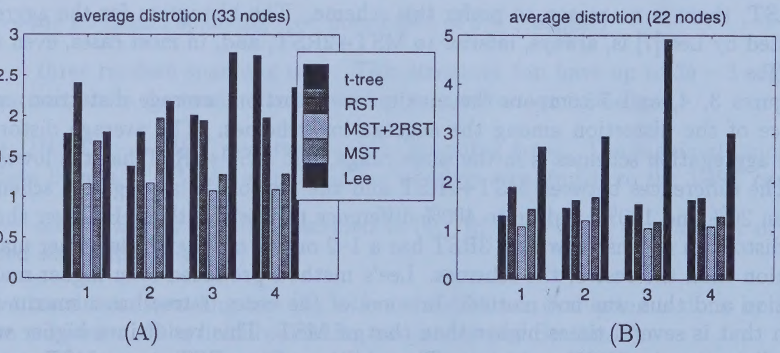


Figure 2: A comparison between the histogram of the aggregation schemes for the 33 border node graph.

aggregation is used. MST was found to perform very well with little to no difference from the full knowledge case. This corresponds to our results that show that MST has low distortion.

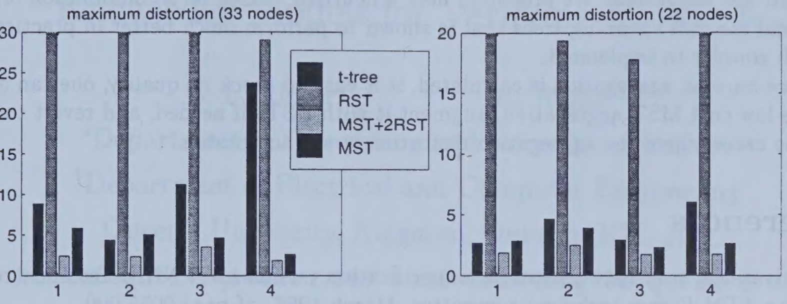
Awerbuch et al. [3] found that the RST aggregation performance is sensitive to the topology, sometimes it is much worse than MST and sometime almost identical. This might be explained by the large variance in the distortion found in this study.

MST+2RST is shown here to be superior to MST. However, in [3], MST is shown to perform very well. It might be due to the simulation of topologies with a small number of border nodes.



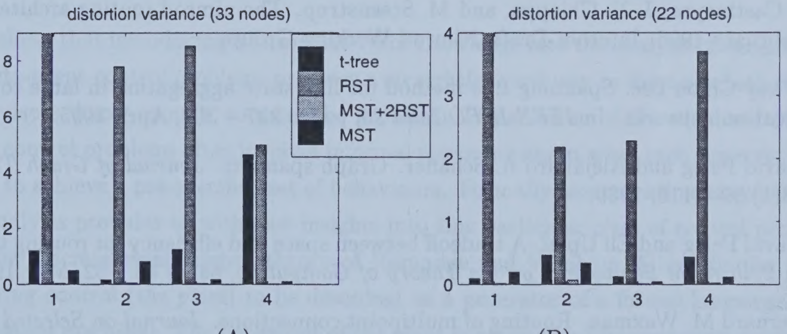
(A) (B)

Figure 3: A comparison between the average distortion of the aggregation schemes. (A) For the 33 border node graph. (B) For the 22 border node graph.



(A) (B)

Figure 4: A comparison between the maximum distortion of the aggregation schemes. (A) For the 33 border node graph. For data sets 1, 2, and 3 the maximum distortion for 3RST is 249, 54 and 50, correspondingly. (B) For the 22 border node graph. For data sets 1 and 4 the maximum distortion for 3RST is 58.5 and 24.2, correspondingly.



(A) (B)

Figure 5: A comparison between the variance of the distortion of the aggregation schemes. (A) For the 33 border node graph. For data set 1 the variance of the distortion for 3RST is 122. (B) For the 22 border node graph. For data set 1 the variance of the distortion for 3RST is 16.2.

5 Concluding Remarks

In the past a theoretically bounded aggregation scheme that is suited for ATM PNNI standard was suggested. We presented here a heuristic, based on a combination of minimum and random spanning trees that is shown to perform much better in practice, and is much simpler to implement.

Since once an aggregation is calculated, it is easy to check its quality, one can always use the low cost MST aggregation, augment it with RSTs if needed, and revert to t-tree in those cases where the aggregation distortion is not acceptable.

References

[1] Private network-network interface specification version 1.0 (PNNI). Technical report, The ATM Forum technical committee, March 1996. af-pnni-0055.000.

[2] I. Althofer, G. Das, D. Dopkin, D. Joseph, and J. Soares. On sparse spanners of weighted graphs. *Discrete and Computational Geometry*, 9:81 – 100, 1993.

[3] Baruch Awerbuch, Yi Du, Bilal Khan, and Yuval Shavitt. Routing through networks with hierarchical topology aggregation. *Journal of High-Speed Networks*, 7(1), 1998.

[4] Baruch Awerbuch and Yuval Shavitt. Topology aggregation for directed graphs. In *Third IEEE Symposium on Computers and Communications*, pages 47 – 52, June 1998.

[5] Yair Bartal. Probabilistic approximation of metric space and its algorithmic applications. In *37th Annual IEEE Symposium on Foundations of Computer Science*, October 1996.

[6] I. Castineyra, J. N. Chiappa, and M. Steenstrup. The nimrod routing architecture, February 1996. Internet Draft, Nimrod Working Group.

[7] Whay Chiou Lee. Spanning tree method for link state aggregation in large communication networks. In *IEEE INFOCOM'95*, pages 297 – 302, April 1995.

[8] David Peleg and Alejandro A. Schäffer. Graph spanners. *Journal of Graph Theory*, 13(1):99 – 116, 1989.

[9] David Peleg and Eli Upfal. A tradeoff between space and efficiency for routing tables. In *20th ACM Symposium on the Theory of Computing*, pages 43 – 52, May 1988.

[10] Bernard M. Waxman. Routing of multipoint connections. *Journal on Selected Areas in Communications*, 6:1617 – 1622, 1988.

Knowledge Models for Verifying Protocols

S. L. Ricker* K. Rudie†*

*Department of Computing and Information Science
†Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario K7L 3N6
ricker@qucis.queensu.ca, rudie@eleceng.ee.queensu.ca

Abstract

There exists a class of decentralized discrete-event control problems where a control solution could be realized if additional events were incorporated into the description of the plant to disambiguate undesirable behaviour from legal behaviour. A state-based knowledge model for the analysis of such control problems is introduced. A communication protocol is used to motivate a possible methodology for identifying states in the plant where the introduction of additional events would produce a control solution.

1 Introduction

We believe that introducing a formal model of knowledge into the analysis of decentralized discrete-event control problems provides a straightforward way to discuss whether or not supervisors 'know enough' to solve the problem. The solution of decentralized discrete-event control problems often involves informal reasoning about what each supervisor must know to achieve a pre-specified set of behaviours. Formally incorporating knowledge into the analysis provides us with new insights into this particular class of control problems.

The discrete-event control theory of Ramadge and Wonham [5] entails the system requiring control (the *plant*) to be described as a generator of a formal language (i.e., a finite-state machine). The behaviour of the plant is represented by sequences constructed from a non-empty set of symbols called an *alphabet*. The alphabet represents the set of all possible *events* that can occur within the system. Transitions from one system state to another do not depend on the passage of time, but rather, on the occurrence of an event. The goal is to develop a control strategy for an overseer, or *supervisor*, that will restrain the behaviour of the plant to that of a pre-specified sublanguage (the *legal language*). The supervisor averts undesirable behaviour of the plant by either preventing some events from taking place or allowing—but not forcing—others to occur.

There are some systems where not all events can be seen by the supervisor. A supervisor thus has only a partial view of the system. When the application requires more than one such supervisor to achieve the desirable behaviour, we have a decentralized discrete-event problem.

Decentralized control problems arise naturally in distributed systems. For instance, in telecommunication problems a sender and receiver are physically separated by a communication channel, so that neither participant sees everything that occurs. Communication protocols—in particular, variations of the alternating bit protocol—have been cast as decentralized discrete-event problems (e.g.,[1, 7, 8, 10]).

In this paper we explore solutions to a particular class of decentralized discrete-event control problems: systems where a control solution could be reached if additional events are added to the alphabet to disambiguate good behaviour from illegal behaviour. We use the verification of communication protocols, in particular, a variation of the alternating bit protocol (following the formulation of the problem in [8]), as a motivating example with which we illustrate an updated version of a knowledge model, initially presented in [6]. Using this model, we identify the family of sequences that cause the protocol to fail. Further, an analysis of these sequences offers possibilities as to when the incorporation of additional events could result in a decentralized control strategy.

2 The Knowledge Model

Our knowledge model is based on the possible-worlds model introduced by Halpern and Moses [3] for analyzing the behaviour of distributed systems. Control decisions in the knowledge model introduced in [6] are made by an *agent* on the basis of a partial view of *sequences* of events generated by the plant. One of the disadvantages of this approach is that the semantics of the knowledge model must be expressed as a finite-state structure, where a state in the knowledge model corresponds to a sequence generated by the plant. In the knowledge model presented here, agents monitor the possible set of *states* the plant could be in.

We illustrate the knowledge model by example. Figure 1(a) contains a finite-state machine that describes the behaviour of a system we want to control. In our example, the plant can perform three events: *a, b, c*. All three events are *controllable*—that is, it is possible to turn these events on or off when necessary. For purposes of this example, we will assume that the plant will be controlled by two agents. In this case, we want a control strategy whereby at least one of the agents prevents event *c* from taking place at plant state 4. The nature of the decentralized problem is such that neither agent sees the entire behaviour of the plant. Agent 1 can only observe events *a* and *c* and is also able to control both these events. Agent 2 observes events *b* and *c* and also controls both of the events it sees.

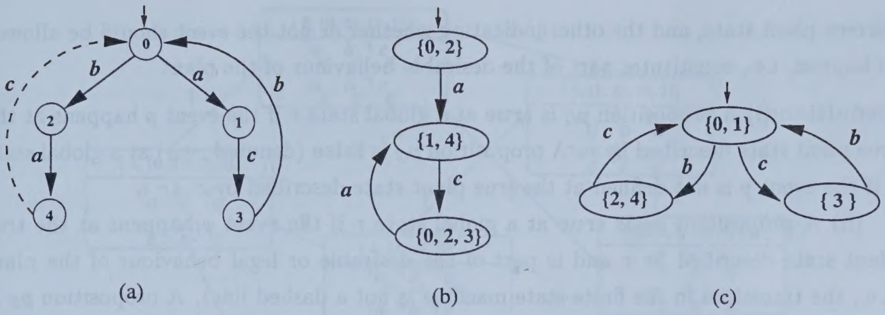


Figure 1: (a) An example plant. Illegal behaviour is designated by a dashed line. (b)The view of the plant held by Agent 1. (c) The view of the plant held by Agent 2. The initial state for each finite-state machine is marked with the small entry arrow.

The partial view of the plant for each agent is shown in figure 1(b) and (c). For example, because Agent 1 does not see event b happening, if it initially sees "nothing", then it does not know if the plant is in state 0 or state 2. If no events have yet occurred, then the plant will indeed be in state 0; however, if b happens, Agent 1 still sees nothing—despite the fact that the plant is really in state 2. Once Agent 1 observes its first occurrence of event a, it still does not know the true plant state. Assume the plant was previously in state 0 and now a happens. The plant is currently at state 1. But Agent 1 considers it possible that the plant could have been at state 2 before a happened. This uncertainty is reflected in the transition from state $\{0, 2\}$ to state $\{1, 4\}$ via event a in figure 1(b).

From these three finite-state machines, we can construct a larger structure that monitors not only the state changes in the plant, but also the partial views each agent has of these state changes. For example, if the plant generates the sequence ac, the plant is at state 3. Agent 1 also sees sequence ac but is unsure if this represents the plant sequences ac, acb, bac or $acbb$. Using figure 1(b), we trace the sequence that Agent 1 sees and arrive at its *local state* $\{0, 2, 3\}$. Agent 2 sees only c occur but there is no uncertainty as to what sequence the plant has just generated. The only scenario under which Agent 2 sees c is if ac is the current plant sequence. Tracing the sequence Agent 2 sees in figure 1(c), we arrive at its local state of $\{3\}$. We combine all this information to build a *global state* in the monitoring structure: $(3, \{0, 2, 3\}, \{3\})$. A global state is an $(n + 1)$-tuple (where n is the number of agents for the system) containing the true plant state and the local state of each agent.

Associated with each global state is a set of logical *propositions*, on which we test knowledge of the occurrence of events. At each global state, every event is represented by two propositions. The first one indicates whether or not the event is defined at the current plant state, and the other indicating whether or not the event should be allowed to happen, i.e., constitutes part of the desirable behaviour of the plant.

Definition:(i) A proposition p_G is **true** at a global state r if the event p happens at the true plant state described by r. A proposition p_G is **false** (denoted $\neg p_G$) at a global state r if the event p is not defined at the true plant state described by r.

(ii) A proposition p_E is **true** at a global state r if the event p happens at the true plant state described by r and is part of the desirable or legal behaviour of the plant (i.e., the transition in the finite-state machine is not a dashed line). A proposition p_E is **false** (denoted $\neg p_E$) at a global state r if either $p_G =$ **false** or if the event p is part of the illegal behaviour of the plant (i.e., the transition is represented by a dashed line).

(iii) We say that the proposition p_G is "event p can occur" and p_E is "event p is legal".

For example, at global state $(3, \{0, 2, 3\}, \{3\})$ we define the following set of propositions: $\neg a_G, \neg a_E, b_G, b_E, \neg c_G, \neg c_E$. Note that at plant state 3 in figure 1(a) only one event is defined, namely event b. Thus we set b_G to **true**. We also set b_E to **true** because event b is part of the legal language. Both a_G and c_G are set to **false** since there are no transitions of either event a or event c at plant state 3. By definition (above), it follows that a_E and c_E are also set to **false**.

The *possible worlds* of each agent are determined by the value of the local states. Two worlds (global states, in our case) "look alike" or are indistinguishable to an agent if their local states are equal. The complete set of possible worlds is captured in a diagram called a *Kripke structure* (see [2] for more details). The Kripke structure for our example is shown in figure 2. Note that each global state has associated with it a set of propositions. States that look alike to agent i are connected by an edge and labeled with an i. For instance, states $(2, \{0, 2\}, \{2, 4\})$, $(2, \{0, 2, 3\}, \{2, 4\})$ and $(4, \{1, 4\}, \{2, 4\})$ are all connected by edges labeled with a 2. These global states are indistinguishable to Agent 2 because its local state, $\{2, 4\}$, is the same in all three states.

An agent has knowledge of a proposition if, at all states it considers indistinguishable, the truth value of the proposition is the same. For example, in figure 2, when the local state of Agent 1 is $\{0, 2\}$, we say that Agent 1 knows that event a should be allowed to happen because propositions a_G and a_E are both true at global states $(0, \{0, 2\}, \{0, 1\})$ and $(2, \{0, 2\}, \{2, 4\})$. More formally, the notation $K_i p$ is interpreted as "agent i knows p". Thus we can state, for example, that at global state $(0, \{0, 2\}, \{0, 1\})$ Agent 1 knows a_G and a_E: $K_1(a_G \wedge a_E)$.

A control solution to the problem exists if, at those states where an event must be prevented from occurring, at least one agent that can control the event has enough knowledge to stop the event. At every global state (and for every event in the alphabet) we want to make certain that one of the following conditions holds: either $\neg p_G$ holds

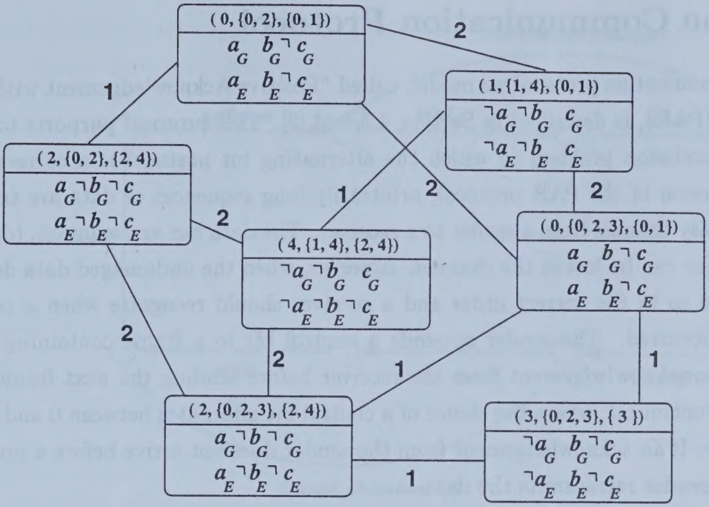

(0, {0, 2}, {0, 1})
$a_G\ b_G\ \neg c_G$
$a_E\ b_E\ \neg c_E$
2
1
(1, {1, 4}, {0, 1})
$\neg a_G\ \neg b_G\ c_G$
$\neg a_E\ \neg b_E\ c_E$
(2, {0, 2}, {2, 4})
$a_G\ \neg b_G\ \neg c_G$
$a_E\ \neg b_E\ \neg c_E$
1
2
2
2
(4, {1, 4}, {2, 4})
$\neg a_G\ b_G\ c_G$
$\neg a_E\ \neg b_E\ \neg c_E$
(0, {0, 2, 3}, {0, 1})
$a_G\ b_G\ \neg c_G$
$a_E\ b_E\ \neg c_E$
2
2
1
1
(2, {0, 2, 3}, {2, 4})
$a_G\ \neg b_G\ \neg c_G$
$a_E\ \neg b_E\ \neg c_E$
(3, {0, 2, 3}, { 3 })
$\neg a_G\ b_G\ \neg c_G$
$\neg a_E\ b_E\ \neg c_E$
1

Figure 2: The Kripke structure for the example in figure 1.

(i.e., event p is not defined in the plant at that state) or p_E holds (i.e., event p is part of the legal language) or there exists an agent i that controls event p such that $K_i \neg p_E$. A system satisfying this criterion is said to be "Kripke-observable".

In our example, we want to make sure that at least one agent has the knowledge to turn off event c at plant state 4. We must check the Kripke-observability conditions for event c at global state $(4, \{1, 4\}, \{2, 4\})$. The first condition is not satisfied since $c_G = \textbf{true}$. The second condition is not satisfied either, because $c_E = \textbf{false}$. Thus we must determine if either agent (since both control event c) has the knowledge to ensure c does not happen. At this global state, Agent 1's local state is $\{1, 4\}$; however, at all global states that share this local state, it is not the case that $K_1 \neg c_E$: at global state $(1, \{1, 4\}, \{0, 1\})$, $c_E = \textbf{true}$. Therefore Agent 1 does not have sufficient knowledge to stop event c from happening. At all global states where Agent 2's local state is $\{2, 4\}$, it is the case that $\neg c_E$ holds. Thus Agent 2 "knows" enough to recommend that event c be disabled at plant state 4.

A *knowledge protocol* or control strategy is constructed as follows: when at least one agent i "knows" $\neg p_E$ at a global state r (for any event p in the system's alphabet that Agent i controls), then a "disable p" action is taken. Otherwise an "enable p" action is issued for that global state. The protocol solves the problem if the set of control actions ensures that only legal sequences and all legal sequences are allowed to occur. The knowledge protocol for our example issues a lone disablement action: at global state $(4, \{1, 4\}, \{2, 4\})$, the control action is "disable c". Thus the protocol admits exactly the legal language of the plant in figure 1.

3 The Communication Protocol

The communication protocol we model, called "Positive Acknowledgement with Retransmission" (PAR), is described in Section 4.3.3 of [9]. This protocol purports to solve the data transmission problem to which the alternating bit protocol is a correct solution. In this version of the PAR protocol, arbitrarily long sequences of data are transmitted along a noisy channel from a sender to a receiver. The data can arrive intact, (detectably) damaged, or can be lost in the channel. However, when the undamaged data does arrive, it must do so in the correct order and a receiver should recognize when a retransmission has occurred. The sender appends a control bit to a frame containing new data and awaits acknowledgement from the receiver before sending the next frame with the attached control bit, where the choice of a control bit alternates between 0 and 1 for each new frame. If an acknowledgement from the sender does not arrive before a pre-specified time, the sender retransmits the data.

The protocol is described over an asynchronous simplex (one way) channel with acknowledgement. We assume that while the channel is used to transmit the frames, the sender and receiver each has access to a buffer (and for illustrative purposes, the buffer has a capacity of one). Thus, messages must be handled in the order in which they are received. Messages arriving at a full buffer are assumed to be lost; however, typically the buffers would be modeled as an infinite first-in first-out (FIFO) queue.

The PAR protocol fails for the following scenario: the sender transmits some data and starts a timer. An acknowledgement does not arrive from the receiver before the timer goes off so the sender assumes the data was lost in the channel and retransmits the data. In the meantime, the first frame of data arrives at the receiver and an acknowledgement is sent. When the acknowledgement reaches the sender, the sender mistakenly assumes that the acknowledgement just received is acknowledging the retransmitted data. A new frame of data is sent to the receiver and is subsequently lost in the channel. The second acknowledgement, for the retransmitted frame, now reaches the sender and is assumed to be for the frame that was (unbeknownst to the sender) actually lost. No retransmission of the lost data will occur and the protocol fails. The classic solution requires that the acknowledgements be indexed by the control bit attached to the data frame.

4 A Knowledge Model of PAR

The knowledge model for the PAR protocol uses the finite-state machines for the sender and the receiver in [8] and the model of the channel from [10], as illustrated in figure 3. The plant is the synchronous product (see [4] for a definition) of these three finite-state machines.

The sender (Agent 1) observes and controls events *getframe*, $send_0$, $send_1$ and *timer*

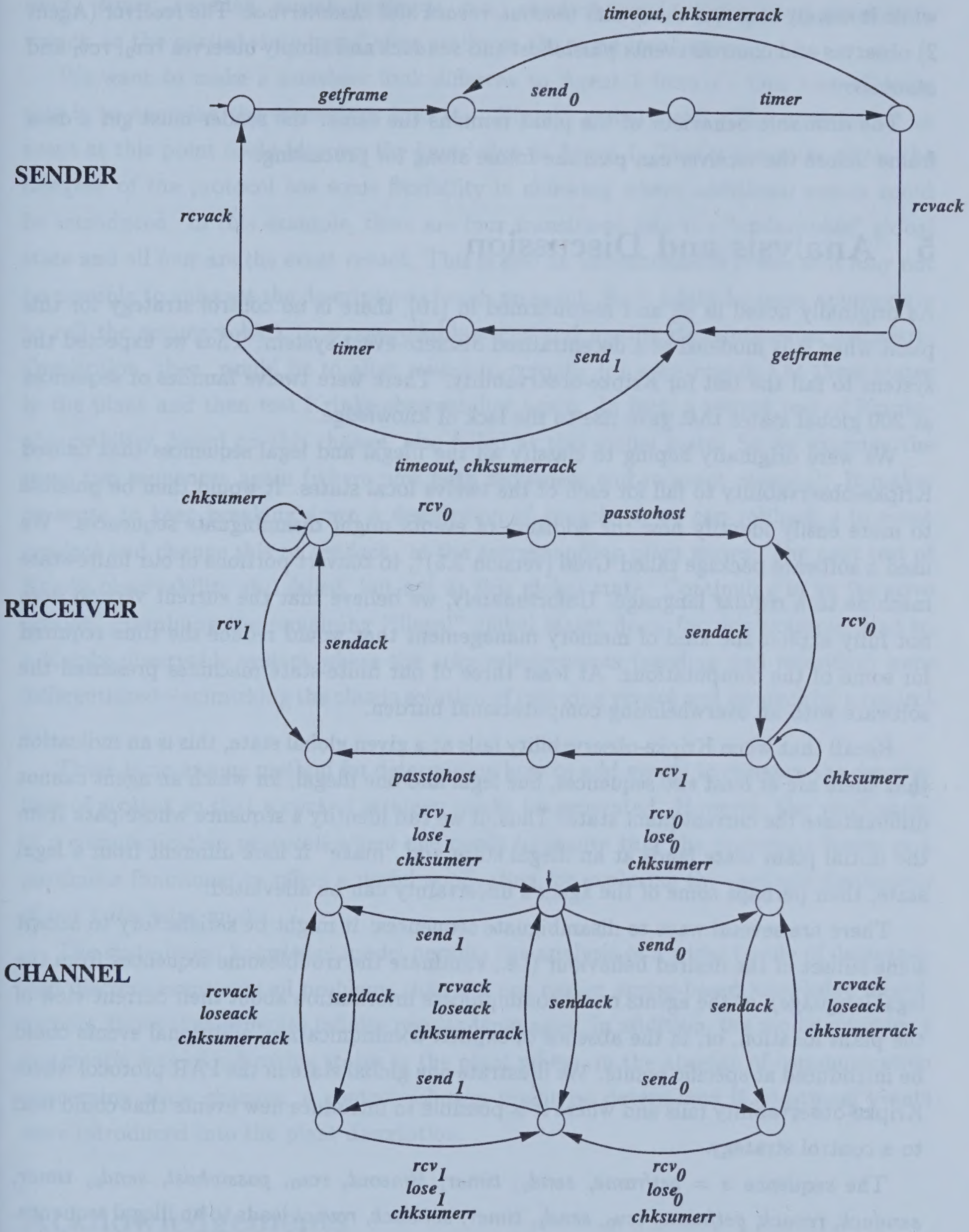

SENDER
getframe
$send_0$
timer
timeout, chksumerrack
rcvack
rcvack
timer
$send_1$
getframe
timeout, chksumerrack
RECEIVER
chksumerr
rcv_0
passtohost
rcv_1
sendack
sendack
rcv_0
passtohost
rcv_1
chksumerr
CHANNEL
rcv_1
$lose_1$
chksumerr
rcv_0
$lose_0$
chksumerr
$send_1$
$send_0$
rcvack
loseack
chksumerrack
sendack
rcvack
loseack
chksumerrack
sendack
sendack
rcvack
loseack
chksumerrack
$send_1$
$send_0$
rcv_1
$lose_1$
chksumerr
rcv_0
$lose_0$
chksumerr

Figure 3: The finite-state machines for the PAR protocol plant.

while it merely observes the events *timeout*, *rcvack* and *cksumerrack*. The receiver (Agent 2) observes and controls events *passtohost* and *sendack* and simply observes rcv_0, rcv_1 and *cksumerr*.

The desirable behaviour of the plant remains the same: the sender must get a data frame before the receiver can pass the frame along for processing.

5 Analysis and Discussion

As originally noted in [8] and re-confirmed in [10], there is no control strategy for this plant when it is modeled as a decentralized discrete-event system. Thus we expected the system to fail the test for Kripke-observability. There were twelve families of sequences at 200 global states that gave rise to the lack of knowledge.

We were originally hoping to classify all the illegal and legal sequences that caused Kripke-observability to fail for each of the twelve local states. It would then be possible to more easily identify how the addition of events might disambiguate sequences. We used a software package called *Grail* (version 2.5)[1], to convert portions of our finite-state machine to a regular language. Unfortunately, we believe that the current version does not fully exploit the kind of memory management that would reduce the time required for some of the computations. At least three of our finite-state machines presented the software with an overwhelming computational burden.

Recall that when Kripke-observability fails at a given global state, this is an indication that there are at least two sequences, one legal and one illegal, for which an agent cannot differentiate the current plant state. Thus, if we can identify a sequence whose path from the initial plant state lands at an illegal state and "make" it look different from a legal state, then perhaps some of the agent's uncertainty can be alleviated.

There are several ways to disambiguate sequences: it might be satisfactory to accept some subset of the desired behaviour (i.e., eliminate the troublesome sequences from the legal language), or the agents could communicate information about their current view of the plant location, or, in the absence of explicit communication, additional events could be introduced at specific points. We illustrate one global state in the PAR protocol where Kripke-observability fails and where it is possible to introduce new events that could lead to a control strategy.

The sequence s = *getframe*, $send_0$, *timer*, *timeout*, rcv_0, *passtohost*, $send_0$, *timer*, *sendack*, *rcvack*, *getframe*, rcv_0, $send_1$, *timer*, *sendack*, *rcvack* leads to an illegal sequence. One of the next events that can happen in the plant is *getframe*, yet no *passtohost* precedes it—a violation of our desired behaviour. As far as Agent 1 is concerned, s is indistinguishable from sequence s' = *getframe*, $send_0$, *timer*, *timeout*, rcv_0, *passtohost*,

---

[1] Available at `http://www.csd.uwo.ca/research/grail`

$send_0$, $timer$, $sendack$, $rcvack$, $getframe$, rcv_0, $sendack$, $send_1$, $timer$, rcv_1, $passtohost$, *rcvack*, as the partial state-based view produces the same local state.

We want to make s somehow look different to Agent 1 from s'. One way of doing this is to examine the transitions into the "illegal" global state. The addition of an event at this point could improve the knowledge of Agent 1. This is the point where the designer of the protocol has some flexibility in choosing where additional events could be introduced. In this example, there are four transitions into the "undesirable" global state and all four are the event *rcvack*. This is also an uncontrollable event, so it may not be possible to enhance the description of such an event. So it might be more appropriate to roll the sequence back to a controllable event and modify descriptions as necessary. One option, then, would be to alter *rcvack* to $rcvack_B$ (or even $rcvack_1$) at these states in the plant and then test Kripke-observability again. In fact, a second test of Kripke-observability, based on this change, also failed at this global state. So we examine the same two sequences again (where now both sequences end in event $rcvack_B$). It makes no sense to keep breaking down a description of *rcvack*, so we can rollback s to event *sendack* and change this to $sendack_A$ at the corresponding plant states. The next test of Kripke-observability still failed, but not at this global state. Continuing in an iterative fashion, examining the remaining "illegal" global states does, for this example, lead to a Kripke-observable system where the acknowledgements (sending and receiving) were differentiated—mimicking the classic solution of indexing *rcvack* and *sendack* by a control bit.

There is no unique method for determining how to add events to enhance the description of a plant so that a control strategy might be generated. However, the verification of a communication protocol, where one needs to ensure that the protocol adheres to a particular functionality, offers a useful application for exploring the uses and limitations of our knowledge model.

The state-based knowledge model permits the analysis of a wider family of decentralized discrete-event control problems than did our earlier *string*-based knowledge model, namely, those that generate infinite regular languages. In addition, the model provides a systematic way of indicating states in the plant where, in the absence of communication concerning state changes, a control solution might be determined if additional events were introduced into the plant description.

Acknowledgements

This work is partially supported by NSERC grant OGP0138887. Thanks also to Gary Anderson, Reza Kopaee and Tim Wright for technical support.

References

[1] R. Cieslak, C. Desclaux, A. S. Fawaz, and P. Varaiya. Supervisory control of discrete-event processes with partial observations. *IEEE Trans. on Automatic Control*, 33(3):249–260, March 1988.

[2] R. Fagin, J.Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge.* MIT Press, 1995.

[3] J.Y. Halpern and Y. Moses. Knowledge and common knowledge in a distributed environment. In *Proc. 3rd ACM Symp. on Principles of Distrib. Comput.*, pages 50–61, 1984.

[4] C. A. R. Hoare. *Communicating Sequential Processes.* Prentice-Hall, Inc., 1985.

[5] P.J. Ramadge and W.M. Wonham. Supervisory control of a class of discrete event processes. *SIAM J. Contr. Optim.*, 25(1):206–230, 1987.

[6] S. L. Ricker and K. Rudie. Know means no: Incorporating knowledge into decentralized discrete-event control. In *Proceedings of the 1997 American Control Conference*, pages 2348–2353, 1997.

[7] K. Rudie and W. M. Wonham. Supervisory control of communicating processes. In H. Ural, Probert R. L., and L. Logrippo, editors, *Protocol Specification, Testing and Verification, X*, pages 243–257. Elsevier Science Publishers B. V., 1990.

[8] K. Rudie and W.M. Wonham. Protocol verification using discrete-event systems. In *Proceedings of the 31st Conference on Decision and Control*, pages 3770–3777, 1992.

[9] A. S. Tanenbaum. *Computer Networks.* Prentice-Hall, Inc., Englewood-Cliffs, NJ, second edition, 1988.

[10] K.C. Wong and J.H. van Schuppen. Decentralized supervisory control of discrete-event systems with communication. In *Proc. WODES'96*, pages 284–289, 1996.

On Fault-Tolerant Routing in Iterated Line Digraph Networks using Optical Interconnections

Abhijit Sengupta[1]
Department of Computer Science
University of South Carolina
Columbia, SC 29208
gupta@cs.sc.edu

Abstract

This paper considers the fault-tolerant routing problem in interconnection networks using wavelength division multiplexed (WDM) all-optical communication. It is known that large network digraphs can be designed from small networks through line digraph iteration. This paper shows how the fault-tolerant routing techniques for small networks can be used for larger networks keeping good fault-tolerant properties in an all-optical network using WDM methods.

I. Introduction

In an optical network, the end-nodes communicate with each other by establishing lightpath(s) between them along optical fiber(s). In all-optical networks, a lightpath between two end-nodes is a path between them through one or several routers using a particular wavelength for each segment of the fiber traversed. The WDM (wavelength division multiplexing) scheme uses frequency multiplexing in the optical domain where several communication channels operate at different carrier frequencies on a single fiber. Unless wavelength conversion can be accomplished in optical domain at a router node, the same wavelength must be maintained along an entire lightpath. A solution to the routing problem in such networks has to address the problem of finding a lightpath between every source-destination end-node pair and the wavelength to be used in the path. A routing scheme is called a static routing scheme if, for every source-destination end-node pair, the routing path and the wavelength used for communication is predetermined and does not change or depend on the communication in progress in the network at any time. Several static routing schemes have been proposed for defining the routing path and the wavelength assignment [18-23] with objectives to minimize the congestion and/or the communication delay, or to maximize the throughput, or to optimize other network performance characteristics. Depending on the routing algorithm, the router nodes are programmed to establish the lightpaths accordingly. In the presence of faults occurring to nodes, links or channels, however, one or more lightpaths may become unusable. Fault avoidance clearly would need the choice of an alternate path and/or channel. For a network using static routing scheme, this will, in turn, need alternate routing information built into the routers for all possible fault occurrences. Moreover, the information indicating which network components are faulty should be available to the routers. In a large network, where several faults can occur, having these information available to all the routers is impractical. Fault-tolerant routing has not been considered in detail except for rings where at most a single fault can be tolerated and alternate routing can be statically determined. In a multihop network, the communication between two nodes go through one or several other intermediate nodes. Depending on the traffic in the network, an intermediate nodes may have to buffer messages before it can forward to the destination node

---

[1] The research was supported in part by a grant from the Office of Naval Research by contract number N00014-97-1-0806.

needing conversion to electronic domain and reconversion to optical domain and such buffering causes significant loss of bandwidth.

In recent years, there have been significant [14-17] research interest on the usage of optical communication between the processors in a multiprocessor system. Considering point-to-point communication between the processors, the logical interconnection between the processors is the same as the physical interconnection between them in a conventional electronic communication system. However, in a network using optical communication, a logical interconnection can be implemented in various physical topologies by appropriate embedding. For example, using a passive star physical topology, two processors communicate by tuning their optical transmitters and receivers to the same wavelengths. The sending processor sends out its message to the star at a particular wavelength and the receiving processor picks up the message tuning the receiver to the same wavelength. The logical interconnection between the different processors can be changed, if necessary, by simply retuning the transmitter and/or receiver wavelength, without any change in the physical interconnection.

A static or fixed routing ρ in a digraph G, representing the logical interconnection, is a function that gives for every ordered pair (x, y) of nodes of G, a directed path $\rho(x, y)$ (represented by a sequence of nodes or directed edges) called the route from x to y. According to the fixed routing strategy, any two nodes communicate with each other using the route between them and using the wavelength channel assigned to it. Clearly, if two routes do not share any edge, the same wavelength can be assigned to them, whereas, if they do, different wavelengths have to be assigned.

The routing problem becomes even more difficult in presence of faults. An edge failure can occur because of transmitter/receiver failure or cut in fiber. A node failure can occur because of failure in processor or optical switching. The idea of using fixed route was considered first in [2], though in a different context. Several studies [3-9] have been reported on the fixed routing strategy. If any node or edge in a route is faulty then the route becomes unusable. In that case, two processors x and y can still communicate with each other using other routes which do not have any faulty node or edge. For example, for some fault-free node α, if $\rho(x, \alpha)$ and $\rho(\alpha, y)$ do not have any faulty node or any faulty edge then $\rho(x, \alpha)$ followed by $\rho(\alpha, y)$ might be used for communication from x to y. Note that, depending on the wavelength channel assignment, wavelength conversion might be needed at the node α.

Given a network digraph G, a routing function ρ and a set of faults F (assume F_v is the set of faulty nodes and F_e is the set of faulty edges) the survival route graph $R(G, \rho)/F$ is defined [2] as a graph with $V - F_v$ as the node set and for $x, y \in V - F_v$, (x, y) is an edge if and only if the route from x to y is fault-free (that is, does not have any faulty node or edge in the route). It is clear that the diameter of the survival route graph gives an upper bound on the number of wavelength conversions that might be needed for using fixed routing with F as the fault set. If wavelength conversion is not used, it gives an upper bound on the number of electronic buffering might be needed for the same fixed routing. Hence, for an efficient performnace, this diameter should be low. Several researchers have suggested different routing functions [2-9] to achieve low diameter of the survival route graphs. Several of these works [2-5] considered undirected graphs as the logical topology which can be used for optical interconnection if every undirected edge is replaced by two oppositely directed edges. Interconnection networks given by digraphs have been considered in [6-9] and properties of such network have been derived to insure constant diameter of survival route graphs. Several other properties of $R(G, \rho)/F$ have been studied elsewhere, an excellent list of references is available in [10]. node connectivity of the network.

It has been shown in [11] that starting from a small digraph, large network topologies can be designed by line digraph iteration. Such construction maintains the same node connectivity and increases the diameter by a minimum amount. It is natural to expect that if an efficient routing algorithm is known for the small digraph that the iteration started with, the routing algorithm should be extendable for larger digraphs as iteration progresses. We will refer to a routing algorithm as an efficient routing algorithm if such a routing algorithm results in a constant value of the diameter of the survival route graph. In this paper, we address the issue of designing efficient routing algorithm so that line digraph iteration does not change the diameter of the survival route graph. For such an algorithm, the number of wavelength conversions or opto-electronic conversion needed is independent of the size of the network and remains independent as the network grows by line digraph iteration.

II. Preliminaries and earlier works

In this section, we present the pertinent definitions and terminologies. Standard graph-theoretic terminologies are not defined and readers are referred to [1] for them. Throughout this paper, the interconnection network will be represented by a digraph $G = (V, E)$ with V as the set of nodes and E as the set of directed edges. An edge from the node u to v will be denoted by (u, v) or by $u \rightarrow v$. Given G, its line digraph $L(G)$ is a digraph where every edge of G is a node of $L(G)$ and if (α, β) and (γ, δ) are edges in G then, there is a directed edge from (α, β) to (γ, δ) in $L(G)$ iff $\beta = \gamma$. For any integer $x > 0$, we will use $L^x(G)$ to denote $L(L^{x-1}(G))$ where $L^0(G) = G$. Let k_G denote the node connectivity of G (we will omit the subscript G when there is no ambiguity). It is well known [10] that $k_{L(G)} \geq k_G$ and the diameter of $L(G)$ can exceed the diameter of G by at most one.

Perhaps the most frequently referred examples of iterated line digraphs are DeBruijn and Kautz digraphs. A DeBruijn digraph, denoted by $DB(r, m)$, is a digraph with r^m nodes defined as follows using an r element alphabet $Z = \{0, 1, 2, \cdots, r-1\}$. Each node is represented by a m-dimensional vector $(a_1 a_2 \cdots a_m)$, where $a_i \in Z, 1 \leq i \leq m$. A node given by $(a_1 a_2 \cdots a_m)$ has an outgoing edge to the node given by the vector $(a_2 a_3 \cdots a_m x)$ where x is any symbol of Z. It has been shown [7, 11] that for any positive integers α, β, $L^\alpha(DB(r, \beta))$ is $DB(r, \alpha + \beta)$. The Kautz digraph, denoted by $K(r, m)$, is a similar digraph with $r^m + r^{m-1}$ nodes. It is defined using an $r + 1$ element alphabet $Z = \{0, 1, 2, \cdots, r\}$. Each node is represented by a m-dimensional vector $(a_1 a_2 \cdots a_m)$, where $a_i \in Z, 1 \leq i \leq m$ and $a_i \neq a_{i-1}, 2 \leq i \leq m$. A node given by $(a_1 a_2 \cdots a_m)$ has an outgoing edge to the node given by the vector $(a_2 a_3 \cdots a_m x)$ where x is any symbol of Z excepting a_m. It has been shown [7, 11] that for any positive integers α, β, $L^\alpha(K(r, \beta))$ is $K(r, \alpha + \beta)$. It has been shown [7] that for DeBruijn or Kautz digraphs, it is simple to design efficient routing algorithm so that the diameter of the survival route graph do not change with line digraph iteration.

Let $F \subset V \cup E$ be the set of faults in the network G where F_v is the set of faulty nodes and F_e is the set of faulty edges. Let t denote the maximum number of faulty components; hence $|F| \leq t$. To have the faulty G remain strongly disconnected, we will assume $t \leq k - 1$. For any node $u \in V$, let $P(u) = \{\alpha | (\alpha, u) \in E\}$ and $S(u) = \{\alpha | (u, \alpha) \in E\}$. $P(u)$ ($S(u)$) denote the set of predecessors (successors) of u in G. Extending this definition, for any $X \subset V$, let $P(X) = \bigcup_{u \in X} P(u) - X$ and $S(X) = \bigcup_{u \in X} S(u) - X$. Let $d_{in}(u)$ and $d_{out}(u)$ respectively denote the indegree and the outdegree of a node u, that is, $d_{in}(u) = |P(u)|$ and $d_{out}(u) = |S(u)|$. Evidently, $d_{in}(u), d_{out}(u) \geq k$ for all u. A routing ρ_G defined for G is a function that gives, for every ordered pair of nodes (x, y), a directed path from x to y in G (if the digraph G is understood, only ρ will be used to denote the routing function. Given a network digraph G, a routing function ρ, a set of faulty nodes F_v and a set of faulty edges F_e, the survival route graph $R(G, \rho)/F$ (where $F = F_v \cup F_e$) is defined [2] as a digraph with $V - F_v$ as the node set such that for $x, y \in V - F_v$, (x, y) is a directed edge if and only if the route $\rho(x, y)$ does not have any faulty node or edge in it. Let $d_G(x, y; F)$ denote the distance from x to y in $R(G, \rho)/F)$. When G and F are understood, we will simply use $d(x, y)$. Let $D(R(G, \rho)/F)$ denote the diameter of the survival route graph $R(G, \rho)/F$.

Given an interconnection network, its fault vulnerability was first studied in [2] in terms of the diameter of its survival route graph and it was shown that for any undirected graph

G, $D(R(G,\rho)/F) \leq max\{2(k-1), 4\}$ and for some special class of graphs, for example, in hypercubes, there exists a shortest path unidirectional routing ρ for which $D(R(G,\rho)/F) = 2$ and also there exists a shortest path bidirectional routing ρ for which $D(R(G,\rho)/F) = 3$. Efficient routing was constructed in [4] in product graphs from the routings in constituent graphs and an upper bound on the diameter of the survival route graph was computed. In [3, 8, 9], efficient routing was constructed to achieve a constant $D(R(G,\rho)/F)$ using the idea of concentrator nodes. A set of nodes with some specific properties were called concentrator nodes and the routing function was defined around them. Several different routing functions were defined in [3]. It was shown that if adequate number of concentrator nodes could be found then ρ can be defined to have $D(R(G,\rho)/F)$ not to exceed 4. Subsequent papers [8, 9] improved this bound to 3. In [5], some sufficient conditions were found for which ρ could be defined to have $D(R(G,\rho)/F)$ not to exceed 3. With a less restrictive condition, routing function was defined to have the diameter not to exceed 5. Digraphs representing interconnection networks were considered in [6-9]. In [7], it has been shown that for a minimum distance routing, DeBruijn and Kautz digraphs have the diameter of the survival route graph at most 2. In [6], the routing techniques of [3] was extended for digraphs to obtain an upper bound of 6 for $D(R(G,\rho)/F)$.

Given a digraph G with n nodes, a large interconnection can be constructed by finding $L^\alpha(G)$ choosing appropriate value of α. If G is a regular digraph with indegree and outdegree d, then $L^\alpha(G)$ is a regular digraph with $d^\alpha n$ nodes and diameter α more than the diameter of G. If some efficient routing ρ is known for G that satisfies some properties, we show that the routing for G can be readily extended for $L^\alpha(G)$ for all α, such that the the diameter of the survival route graph for $L^\alpha(G)$ for all α is the same as that for G.

III. Routing in iterated line digraphs

In this section, before we discuss the routing schemes in iterated line digraphs, we explain the motivation for doing so.

A. Motivation

DeBruijn and Kautz digraphs are canonical examples of iterated line digraphs. In [13], the authors have proposed two classes of digraphs which, as interconnection digraphs, have interesting properties of high connectivity and low diameter. One such class of digraphs is defined as follows. A regular digraph with n nodes and degree d and denoted by $B(n,d)$ has the node numbered $0, 1, \ldots, n-1$. A node numbered i has outgoing edges to nodes numbered $id + x, 0 \leq x \leq d-1$ where multiplications and additions are in modulo n. It is simple to see from the following discussions that these classes of digraphs satisfy the property that $B(nd,d) = L(B(n,d))$. Clearly starting from any digraph in this class, the line digraph iteration produces digraphs in the same class. For a special case, when the starting digraph has d nodes, DeBruijn graphs are produced after each iteration.

Given a $B(n,d)$ with the nodes numbered, number the edges as follows. Assign the number $id + x$ (modulo nd) to the edge $i \rightarrow id + x, 0 \leq x \leq d-1$. The outgoing edges from $id + x$ then get numbered as $id^2 + xd + y$ (modulo nd), $0 \leq y \leq d-1$. In the $L(B(n,d))$, clearly there are nd nodes. Assign each node the number that was assigned to the corresponding edge of $B(n,d)$. It is simple to see that using this numbering the line digraph is a $B(nd,d)$.

The other class of digraphs considered in [13] is defined as follows. A regular digraph with n nodes and degree d has the node numbered $0, 1, \ldots, n-1$. We will denote the digraph by $K(n,d)$. In $K(n,d)$, a node numbered i has outgoing edges to nodes numbered $-id - x, 1 \leq x \leq d$, where multiplications, additions etc. are in modulo n. In following, we show that, $K(nd,d) = L(K(n,d))$. Thus starting from any digraph in this class, the line digraph iteration produces digraphs in the same class. For a special case, when the starting digraph has $d+1$

nodes. Kautz digraphs are produced after each iteration. The edge from the node numbered i to the node numbered $j = -id - x$ (modulo n) will be denoted by $i \xrightarrow{x} j$.

Assign integer numbers to the edges of $K(n,d)$ as follows. For an edge $i \xrightarrow{x} j$, assign the integer number $id + x - 1$ to it. Since $0 \leq i \leq n-1$ and $1 \leq x \leq d$, the highest and lowest integers assigned to the edges are respectively 0 (for $i = 0$ and $x = 1$) and $nd - 1$ (for $i = n-1$ and $x = d$). We claim that all the assigned integers are distinct. If possible, assume that there are two edges $i \xrightarrow{x_1} j$ and $k \xrightarrow{x_2} m$ which are assigned the same integer. In that case, $id + x_1 - 1 = kd + x_2 - 1$, implying $(i-k)d = x_2 - x_1$, which is impossible, since highest possible value of $x_2 - x_1$ is $d - 1$. Since there are nd edges, each edge is assigned a distinct integer. Construct $L(K(n,d))$ and number the nodes of $L(K(n,d))$ same as the corresponding edge of $K(n,d)$ it represents. Consider any walk $i \xrightarrow{x_1} j \xrightarrow{x_2} k$ of length 2 in $K(n,d)$. Hence $j \equiv -i\ d - x_1\ (\mathit{modulo}\ \ n)$, which implies $j + id + x_1 = pn$ for some positive integer p. Also $k \equiv -j\ d - x_2\ (\mathit{modulo}\ \ n)$, which implies $k + jd + x_2 = qn$ for some positive integer q, that is, $k + pnd - id^2 - x_1 d + x_2 = qn$. Note that

$$\begin{aligned} -(id - x_1 - 1)d - (d + 1 - x_2) &= -id^2 - x_1 d - 1 + x_2 \\ &= qn - k - pnd - 1 \\ &= jd + x_2 - 1 - pnd \end{aligned}$$

Hence $-(id - x_1 - 1)d - (d + 1 - x_2) \equiv jd + x_2 - 1\ (\mathit{modulo}\ \ nd)$. This implies that, in $K(nd,d)$, there is an edge $id + x_1 - 1 \xrightarrow{d+1-x_2} jd + x_2 - 1$. However, $id + x_1 - 1$ and $jd + x_2 - 1$ are the integers assigned to the nodes respectively representing the edges $i \xrightarrow{x_1} j$ and $j \xrightarrow{x_2} k$. Hence $K(nd,d) = L(K(n,d))$.

The above results show that very rich class of interconnection network can be designed with line digraph iteration technique. As the network grows in size, designing efficient fault-tolerant routing algorithms become difficult. We show that there exist sufficient conditions which when satisfied can allow us to design routing algorithms for large networks using similar algorithms for smaller networks if larger networks were formed by line digraph iterations.

B. Routing Issue

Given a digraph G and one of its nodes u_i, we define a new digraph G_i, which is obtained by adding a new node u_{i1} to G such that every node having an edge directed to (or from) u_i has a directed edge to (or from) u_{i1}. It is simple to note that if G has node connectivity k then so does G_i. Also, given a routing function ρ defined for G, ρ can be trivially extended to provide a routing for G_i.

Given a digraph G and a routing function ρ defined for it, ρ will be called prefix consistent routing (in short, a pc routing) iff the following conditions hold.

(i): $\rho(x,y)$ is given by the edge (x,y) if $(x,y) \in E$;

(ii): for any pair of nodes x and y in G and for any node z in the path given by $\rho(x,y)$, the subpath from x to z in $\rho(x,y)$ is $\rho(x,z)$.

It is simple to see that a minimum distance pc routing $\rho(x,y)$ always exists in a digraph. For any given node x, such a $\rho(x,y)$ for all y can be constructed by forming a breadth-first search tree routed at x where the paths in the tree define the ρ function. Given a minimum distance pc routing ρ_G in a digraph G, it can be trivially extended for G_i for any i by adding $\rho(\alpha, u_{i1}) = \rho(\alpha, u_i)$ and $\rho(u_{i1}, \alpha) = \rho(u_i, \alpha)$ for all nodes α of G_i. Evidently the extension is a minimum distance pc routing for G_i. Henceforth, whenever we refer to a routing in G_i, we will mean the extension of the corresponding routing in G.

A routing function ρ_G in a k-connected digraph G will be defined to satisfy the property π_{kp}, for a given integer p, if

(a): for every pair of nodes x, y in G and for every fault set F, $|F| \leq k-1$,$d_G(x,y;F) \leq p$ and

(b): in every G_i, for every pair of nodes x, y and for every set of faulty edges F with $|F| = k$ such that $\exists \alpha \in S(x)$ and $\exists \beta \in P(y)$, so that $(x,\alpha), (y,\beta) \in F$, $D_{G_i}(x,y;F) \leq p$.

Given a digraph G, with node connectivity k, assume that F is the set of faults and every element of F is an edge. If $|F| \leq k-1$, it is known that a directed path between any pair of fault-free nodes exists that avoids the faulty edges. For some special node pairs and fault condition, a directed path might exist even for larger faulty edge set. Consider any two distinct nodes x, y of G, such that (x,y) is not an edge and suppose that x has outgoing edges to $a_1, a_2, \cdots, a_p$ and y has incoming edges from $b_1, b_2, \cdots, b_q$. Suppose $|F| \leq k$ (that is, equal to the node connectivity) and F contains (x, a_i) and (b_j, y) for some i and j. Since G has edge connectivity at least k, using the result of [12], k pairwise edge disjoint paths can be constructed from the a's to the b's such that one of them is from a_i to b_j. Thus among the k edge disjoint paths, there is a path with two faulty edges and hence even with k faulty edges, a directed path from x to y can always be found that avoids the faulty edges. Such a fault set will be denoted by F_{xy}. In short, given a digraph G with node connectivity k and two nodes x and y, if (x,y) is not an edge then F_{xy} will denote any set of at most k faulty edges which contains at least one outgoing edge from x and one incoming edge to y. If (x,y) is an edge, then F_{xy} can have at most $k-1$ faulty edges.

Given a minimum distance pc routing ρ_G in a digraph G, we define a routing $\rho_{L(G)}$ in $L(G)$ as follows.

$L1$: $\rho_{L(G)}((x,y),(z,w))$ is given by the edge (x,y) to (z,w) if $y = z$;

$L2$: if $\rho_G(y,z)$ is given by the node sequence $yx_1x_2\cdots x_m z$ in G, then $\rho_{L(G)}((x,y),(z,w))$ is given by the node sequence $(x,y)(y,x_1)(x_1,x_2)\cdots(x_m,z)(z,w)$ in $L(G)$.

It is simple to see that given a pc routing ρ_G in G, the routing $\rho_{L(G)}$ is a pc routing in $L(G)$. The main result of this section is given by the following theorem. We later show that the result of [7] that for a minimum distance routing in DeBruijn or Kautz digraph, the survival route digraph has diameter two follows directly from the following theorem.

Theorem 1 : In a k-connected digraph G, suppose there exists a pc routing ρ such that
(a) for every fault set F, $|F| \leq k-1$, $D(R(G,\rho_G)/F) \leq p$ and
(b) for every pair of nodes (need not be distinct) u_x and u_y in G, $d_{G_y}(u_x, u_{y1}; F_{x,y1}) \leq p$.
Then for every $\alpha \geq 0$, and for every fault set F^+, $|F^+| \leq k-1$, $D(R(L^\alpha(G), \rho_{L^\alpha(G)}/F^+) \leq p$.

Before we formally prove the above theorem, some properties of iterated line digraphs needs to be analyzed. Let $u_1, u_2, \cdots, u_n$ be the nodes of G. Since every dipath of length q in G represents a node in $L^q(G)$, we will denote any node of $L^q(G)$ by a node sequence of G of length $q+1$. It might be noted that any set of faults in $L^q(G)$ can be considered as composed of only node faults of $L^q(G)$ unless the source node has a directed edge to the destination node and that edge is faulty. We will consider these two cases separately. Since every node of $L^q(G)$ represents an edge of $L^{q-1}(G)$ and hence any set of node faults in $L^q(G)$ can be viewed as a set of edge faults in $L^{q-1}(G)$.

We prove the above theorem by induction on α.

Basis step : We show that the theorem is true for $\alpha = 1$. Let u_iu_j and u_xu_y be the source and the destination nodes respectively in $L(G)$. We consider two cases.

Case 1 : $j \neq x$. Thus $u_iu_j \rightarrow u_xu_y$ does not exist in $L(G)$ and the set of faults F^+ in $L(G)$ can be considered to be its node faults. Equivalently all these node faults can be considered as edge faults in G. If $L(G)$ has fewer than k faults then so does G. Let F be the set of equivalent faults in G and let $u_jz_1z_2\cdots z_{m-1}u_x$ be a fault-free dipath in $R(G,\rho)/F$ of length m such that $m \leq p$.

Let us call u_j by z_0 and u_x by z_m for uniformity. Evidently, the dipath $\rho(z_t, z_{t+1})$, $0 \leq t \leq m-1$, is fault-free. Suppose in G, $\rho(z_t, z_{t+1})$, $0 \leq t \leq m-1$, is given by $z_t w_{t,1} w_{t,2} \cdots w_{t,l_t} z_{t+1}$. Then since ρ is a pc routing in G, $\rho(z_t, w_{t,l_t})$, is given by the fault-free path $z_t w_{t,1} w_{t,2} \cdots w_{t,l_t}$. Hence in $L(G)$, according to $L1$ and $L2$, the following fault-free paths given by $\rho_{L(G)}$ exist.

$$\begin{gathered} u_i z_0 \rightarrow z_0 w_{0,1} \rightarrow w_{0,1} w_{0,2} \rightarrow \cdots \rightarrow w_{0,l_0} z_1 \\ w_{0,l_0} z_1 \rightarrow z_1 w_{1,1} \rightarrow w_{1,1} w_{1,2} \rightarrow \cdots \rightarrow w_{1,l_1} z_2 \\ \vdots \\ w_{m-2,l_{m-2}} z_{m-1} \rightarrow z_{m-1} w_{m-1,1} \rightarrow w_{m-1,1} w_{m-1,2} \rightarrow \cdots \rightarrow w_{m-1,l_{m-1}} z_m \rightarrow z_m u_y \end{gathered}$$

Hence in $R(L(G), \rho_{L(G)})/F^+$, there exists a path from $u_i u_j$ to $u_x u_y$ of length $m \leq p$.

Case 2 : $j = x$. We will assume that the edge from $u_i u_x$ to $u_x u_y$ is faulty. Because if it is not, the case is trivial. Let F^- be the set of other faults in $L(G)$. The faults in F^- can be considered to be node faults in $L(G)$ and hence edge faults in G. Consider the digraph G_x and assume that the fault set F' in G_x is given by $F' = F^- \cup \{(u_i, u_{x1}), (u_x, u_y)\}$. (Note the definition of G_x and u_{x1} as given in the beginning of this section.) Thus $|F'| \leq k$ and F' contains at least one outgoing edge from u_y and one incoming edge to u_{y1}. If the condition (b) of Theorem 1 holds then $d_{G_x}(u_x, u_{x1}; F') \leq p$. Let $u_x z_1 z_2 \cdots z_{m-1} u_{x1}$ be a fault-free dipath in $R(G_x, \rho)/F'$ of length m such that $m \leq p$. Using exactly the same argument as in Case 1 (by replacing u_j and u_x of Case 1 by u_x and u_{x1} respectively), we can infer that there exists a path from $u_i u_x$ to $u_x u_y$ of length $m \leq p$ in $R(L(G), \rho_{L(G)})/(F^- \cup (u_i u_x, u_x u_y))$. This completes the basis step of the induction argument.

Induction hypothesis : Let us assume that there exists some fixed t, $t \geq 1$, such that the theorem is true for all $\alpha \leq t$.

Induction step : Let $v_1, v_2, \cdots$ be the nodes of $L^t(G)$. Let (v_i, v_j) and (v_x, v_y) be the source and the destination nodes respectively in $L^{t+1}(G)$. Thus each of the v's is a sequence of u's of length $t+1$. We consider two cases.

Case 1 : $j \neq x$. Thus $(v_i, v_j) \rightarrow (v_x, v_y)$ does not exist in $L^{t+1}(G)$ and the set of faults F^+ in $L^{t+1}(G)$, $|F^+| \leq k$, can be considered to be its node faults and hence can be considered as edge faults in $L^t(G)$. Now using exactly the same argument as in case 1 above (using $L^t(G)$ in place of G), we can infer that there exists a fault-free dipath in $L^{t+1}(G)$ from (v_i, v_j) to (v_x, v_y) of length at most the diameter of the survival route graph of $L^t(G)$.

Case 2 : $j = x$. We will assume that the edge from (v_i, v_x) to (v_x, v_y) is faulty. Because if it is not, the case is trivial. Let F^- be the set of other faults in $L^{t+1}(G)$. Let $v_i = u_{z_1} u_{z_2} \cdots u_{z_{t+1}}$, $v_x = u_{z_2} u_{z_3} \cdots u_{z_{t+2}}$ and $v_y = u_{z_3} u_{z_4} \cdots u_{z_{t+3}}$. For any β, $0 \leq \beta \leq t+1$, let us define

$$X_\beta = \{u_{z_1} u_{z_2} \cdots u_{z_{\beta+1}}, u_{z_2} u_{z_3} \cdots u_{z_{\beta+2}}, \cdots, u_{z_{t-\beta+3}} u_{z_{t-\beta+4}} \cdots u_{z_{t+3}}\}.$$

The faults in F^- can be viewed as node faults by assuming one endpoint of faulty edge to be faulty such that no node of X_{t+1} is faulty. Once the faulty nodes of $L^\beta(G)$ are known, $0 \leq \beta \leq t+1$, each such faulty node can be thought to represent a faulty edge of $L^{\beta-1}(G)$ and can be replaced by a faulty node of $L^{\beta-1}(G)$ such that at least one endpoint of a faulty edge of $L^{\beta-1}(G)$ is faulty and no node of $X_{\beta-1}$ is faulty. It is simple to see such a faulty node exists in $L^{\beta-1}(G)$. Repeating this process, a set of faulty edges F_0 can be found in G. Consider F' to be the set of faulty edges in G where $F' = F_0 \cup \{(u_{z_1}, u_{z_2}), (u_{z_{t+2}}, u_{z_{t+3}})\}$. Assuming the condition (b) of Theorem 1 to hold, $d_{G_{z_2}}(u_{z_{t+2}}, u_{z_2 1}; F') \leq p$. Thus a fault-free path of length at most p exists in $R(G_{z_2}, \rho)/F'$ from $u_{z_{t+2}}$ to $u_{z_2 1}$.

Applying the rules $L1$ and $L2$ exactly the same way as in the case 1 of the basis step, we find that a fault-free path of length at most p exists in the survival graph of $L(G)$ from

$u_{z_{t+1}}u_{z_{t+2}}$ to $u_{z_2}u_{z_3}$. Repeating the application of the rules $L1$ and $L2$ the same way to the fault-free path found in the survival graph of $L^{\gamma}(G)$, $1 \leq \gamma \leq t$, we eventually find the existence of a fault-free path of length at most p in the survival graph of $L^{t+1}(G)$ from $u_{z_1}u_{z_2}\cdots u_{z_{t+2}}$ to $u_{z_2}u_{z_3}\cdots u_{z_{t+3}}$. Then the proof follows by induction. This completes the proof of the Theorem 1.

In the following, we will look into the application of the Theorem 1 in some digraphs. The above theorem when applied to DeBruijn digraph shows that these digraphs have survival graph of diameter at most two if minimum distance routing is used. This is a result which was proven earlier in [7].

It is shown in [7, 11] that the node connectivity of $DB(r,m)$ is $r-1$. It can be easily seen that the minimum distance path between any two nodes in $DB(r,m)$ is unique and the minimum distance routing is also a pc routing. In the following discussion on $DB(r,m)$, we will assume the routing ρ to be the minimum distance routing.

It is trivial to see that so long as the number of faults is fewer than $r-1$, $DB(r,1)$ has a survival route graph of diameter at most two, because $DB(r,1)$ is a digraph in which each node has an edge to every other node. Let us consider $DB(r,2)$.

Since the minimum distance path between any pair of nodes is unique, we will denote by $path(x,y)$ to represent the minimum distance path from x to y. Let $z=(z_1,z_2)$ be the source and $w=(w_1,w_2)$ be the destination node.Let us use the notation $x_i=(z_2,i)$ and $y_i=(i,w_1)$, $i \in Z$.

In the following, we show that the conditions of Theorem 1 hold for $DB(r,2)$ for $p=2$. We will consider several cases.

Case 1 : $z_1 \neq z_2$ and $z_2 \neq w_1$. Consider r paths from z to w such that the i path, $0 \leq i \leq r-1$, is given by z to x_i followed by $path(x_i,w)$. Ignoring the path $z \rightarrow x_{w_1} \rightarrow w$, the remaining $r-1$ paths are node disjoint. Hence the condition (a) of Theorem 1 holds for $p=2$ for $r-2$ faults. For a set of faulty edges F, $|F| \leq r-1$, if $(z,x_{w_1}) \in F$ then one fault-free path exists and hence the condition (b) of theorem 1 holds for $p=2$. Assume $(z,x_{w_1}) \notin F$. The path $z \rightarrow x_{w_1} \rightarrow (w_1,w_1) \rightarrow w$ is fault-free and since $z_2 \neq w_1$, the edge $z \rightarrow (w_1,w_1)$ exists in the survival graph giving $p=2$ and hence the condition (b) of theorem 1 holds for $p=2$.

Case 2 : $z_1 \neq z_2$ and $z_2 = w_1$. Consider r paths from z to w such that one of them is $z \rightarrow w$ and others are z to x_q followed by $path(x_q,w)$ where $0 \leq q \leq r-1$ and $q \neq w_2$. It is simple to check that these paths are node disjoint and hence one of them is fault-free for $|F| \leq r-1$ and the conditions of Theorem 1 hold for $p=2$.

Case 3 : $z_1 = z_2$ and $w_1 \neq w_2$ and $z_2 \neq w_1$. Consider $r-1$ paths from z to w such that the ith path, $0 \leq i \leq r-1$, $i \neq z_i$, is given by z to x_i followed by $path(x_i,w)$. It is simple to check that these paths are node disjoint. For at most $r-2$ faults, there is a fault-free path and hence the condition (a) of Theorem 1 holds for $p=2$. For a set of faulty edges F, $|F| \leq r-1$, if all the faulty edges do not appear in these paths then the condition (b) of Theorem 1 trivially holds for $p=2$. If the edge (z,x_{w_1}) is fault-free then the path $z \rightarrow x_{w_1} \rightarrow (w_1,w_1) \rightarrow w$ is fault-free as the node (w_1,w_1) does not appear in the above $r-1$ paths. Since $z_2 \neq w_1$, the edge $z \rightarrow (w_1,w_1)$ exists in the survival graph giving $p=2$. If the edge (z,x_{w_1}) is faulty, let (z,x_i) and (y_j,w) be two faulty edges in F. Note that the edges (z,x_j) and (y_i,w) are fault-free, otherwise we would have had a fault-free path among those $r-1$ paths. The path $z \rightarrow x_j \rightarrow (j,i) \rightarrow y_i \rightarrow w$ is a fault-free path for which the edges $z \rightarrow (j,i)$ and $(j,i) \rightarrow w$ exist in the survival graph giving $p=2$. Hence the condition (b) of Theorem 1 holds for $p=2$.

Case 4 : $z_1 = z_2$ and $w_1 \neq w_2$ and $z_2 = w_1$. In $DB(r,2)$, the edge $z \rightarrow w$ exists. Assume this edge is faulty otherwise the case is trivial. Since $z \rightarrow w$, $|F| \leq r-2$. Consider $r-1$ paths from z to w such that one of them is $z \rightarrow w$ and others are z to x_q followed by $path(x_q,w)$ where $0 \leq q \leq r-1$ and $q \neq w_1$ and $q \neq z_1$. It is simple to check that these paths are node disjoint and hence one of them is fault-free for $|F| \leq r-2$ and hence the conditions of Theorem 1 hold for $p=2$.

Case 5 : $z_1 = z_2$ and $w_1 = w_2$ and $z_2 \neq w_1$. Consider $r-1$ paths from z to w such that the ith path, $0 \leq i \leq r-1$, $i \neq z_i$, is given by z to x_i followed by $path(x_i, w)$. It is simple to check that these paths are node disjoint. For at most $r-2$ faults, there is a fault-free path and hence the condition (a) of Theorem 1 holds for $p=2$. For a set of faulty edges F, $|F| \leq r-1$, if all the faulty edges do not appear in these paths then the condition (b) of Theorem 1 trivially holds for $p=2$. If the edge (z, x_{w_1}) is fault-free then suppose (z, x_i) is a faulty edge in F. Note that the edge (y_i, w) is fault-free, otherwise we would have had a fault-free path among those $r-1$ paths. The path $z \rightarrow x_{w_1} \rightarrow (w_1, i) \rightarrow w$ is fault-free as the node (w_1, i) does not appear in the above $r-1$ paths. Since $z_2 \neq w_1$ and $i \neq w_1$, the edge $z \rightarrow (w_1, i)$ exists in the survival graph and so does $(w_1, i) \rightarrow w$ giving $p=2$. If the edge (z, x_{w_1}) is faulty, then an argument as in Case 3 infers $p=2$. Hence the condition (b) of Theorem 1 holds for $p=2$.

Case 6 : $z_1 = z_2$ and $w_1 = w_2$ and $z_2 = w_1$. This case can occur since z and w need not be distinct. Consider $r-1$ paths from z to w such that the ith path, $0 \leq i \leq r-1$, $i \neq z_i$, is given by z to x_i followed by $path(x_i, w)$. It is simple to check that these paths are node disjoint. For at most $r-2$ faults, there is a fault-free path and hence the condition (a) of Theorem 1 holds for $p=2$. Let (z, x_i) and (y_j, w) be two edges in F. The path $z \rightarrow x_j \rightarrow (j, i) \rightarrow y_i \rightarrow w$ is fault-free. Since $z_2 \neq j$ and $i \neq w_1$, the edges $z \rightarrow (j, i)$ and $(j, i) \rightarrow w$ exist in the survival graph giving $p=2$. Hence the condition (b) of Theorem 1 holds for $p=2$.

Since $DB(r,2)$ satisfies the conditions of Theorem 1, any DeBruijn digraph $DB(r,m)$, $m \geq 2$, which can be represented as $L^{m-2}(DB(r,2))$ has the diameter of the survival route graph as two whenever the minimum distance routing is used.

Similar results follow for Kautz digraph as well. A Kautz digraph, denoted by $K(r,m)$, is a digraph with $r^m + r^{m-1}$ nodes. It can be defined using an $r+1$ element alphabet $Z = \{0, 1, 2, \cdots, r\}$. Each node is represented by a m-dimensional vector where each component of the vector is a symbol of Z, with the restriction that two consecutive symbols are different. Just like a $DB(r,m)$, a node given by the vector $(a_1, a_2, \cdots, a_m)$ has an outgoing edge to the node given by the vector $(a_2, a_3, \cdots, a_m, x)$ where x is any symbol of Z other than a_m. It has been shown in [7, 11] that $K(r,m)$ is $L^{\alpha}(K(r,\beta))$, $1 \leq \alpha, \beta \leq m-1$, where $\alpha + \beta = m$. It is shown in [7, 11] that the node connectivity of $K(r,m)$ is r. It can be easily seen that the minimum distance path between any two nodes in $K(r,m)$ is unique and the minimum distance routing is also a pc routing. Using similar arguments as in the case of the $DB(r,m)$, we can show that $L^{m-2}(K(r,2))$ (which is $K(r,m)$) has the diameter of the survival route graph as two whenever the minimum distance routing is used.

References

1. C. Berge, Graphs and Hypergraphs, North-Holland, New york, 1973.

2. D. Dolev, J. Halpern, B. Simons and R. Strong, "A new look at fault-tolerant network routing", Information and Computation, vol. 72, pp. 180-196, 1987.

3. D. Peleg and B. Simons, "On fault tolerant routing in general networks", Information and Computation, vol. 74, pp. 33-49, 1987.

4. A. Broder, D. Dolev, M. Fischer and B. Simons, "Efficient fault tolerant routings in networks", Information and Computation, vol. 75, pp. 52-64, 1987.

5. M. Imase and Y. Manabe, "Fault tolerant routings in k-connected networks", Information Processing Letters, vol. 28, pp. 171-175, 1988.

6. Y. Manabe, M. Imase and T. Soneoka, "Reliable and efficient fixed routings on digraphs", Transactions of IEICE, vol. E 71, pp. 1212-1220, 1988.

7. N. Homobono and C. Peyrat, "Fault tolerant routings in Kautz and DeBruijn networks", Discrete Applied Mathematics, vol. 24, pp. 179-186, 1989.

8. A. Sengupta and C. D. Elfe, "On fault-tolerant routing in interconnection digraphs", Proc. 1992 International Conference on Parallel Processing, pp. III-212 - III-215, August, 1992.

9. K. Wada and K. Kawaguchi, "Efficient fault-tolerant fixed routings on $(k+1)$-connected digraphs", Discrete Applied Mathematics, vol. 37/38, pp. 339-352, 1992.

10. J-C. Bermond, N. Homobono and C. Peyrat, "Large fault-tolerant interconnection networks", Graphs and Combinatorics, vol. 5, no. 2, pp. 107-123, 1989.

11. M.A. Fiol, J.L.A. Yebra and I.Alegre, "Line digraph iteration and the (d,k) digraph problem", IEEE Trans. Computers, vol. C-33, pp. 400-403, 1984.

12. W. Mader, "On a property of n-edge-connected digraphs", Combinatorica, vol. 1, no. 4, pp. 385-386, 1981.

13. M. Imase, T. Soneoka and K. Okada, "Connectivity of regular directed graphs with small diameters", IEEE Trans. Computers, vol.C-34, pp.267-273, Mar. 1985.

14. T. Szymanski, "Hypermeshes: optical interconnection networks for parallel computing", Journal of Parallel and Distributed Computing, vol. 26, no. 1, pp. 1-23, 1995.

15. P. Dowd, "High performance interprocessor communication through optical wavelength division multiple access channels", Proc. 18th Annual International Symposium on Computer Architecture, pp. 96-105, 1991.

16. J. Wailes and D. Meyer, "Multiple channel architecture - a new optical interconnection strategy for massively parallel computers", Journal of Lightwave Technology, vol. 9, no. 12, pp. 1702-1716, 1991.

17. Z. Guo, R. Melhem, R. Hall, D. Chiarulli and S. Levitan, "Pipelined communication in optically interconnected arrays", Journal of Parallel and Distributed Computing, vol. 12, pp. 269-282, 1991.

18. R. Ramaswami and K. N. Sivarajan, "Routing and Wavelength assignment in all-optical networks", Tech Rep. RC 19592, IBM Research Report 1994.

19. B. Mukherjee, S. Ramamurthy, D. Banerjee and A. Mukherjee, "Some principles for designing a wide-area optical network", Proc. IEEE INFOCOM 94, 1994.

20. R. Ramaswami and K. N. Sivarajan, "Design of logical topologies for wavelength routed all-optical networks", Proc.IEEE INFOCOM 95, pp 1316-1325, 1995.

21. K. Sivarajan and R. Ramaswami, "Lightwave Networks Based on de Bruijn Graphs," IEEE/ACM Transactions on Networking, vol. 2, No. 1, pp. 70-79, Feb 1994.

22. R. Ramaswami and K. N. Sivarajan, "Optimal Routing and Wavelength Assignment in All-Optical Networks," Proc. IEEE INFOCOM 94, pp. 110-119 Jun. 1994.

23. M. Ajmone Marsan, A. Bianco, E. Leonardi and F. Neri, "Topologies for Wavelength-Routing All-Optical Networks", IEEE/ACM Trans. Networking, vol. 1, Oct. 1993, pp 534–546.

Low-Density Parity Check Codes for Fading Channels with Memory

Andrew P. Worthen* and Wayne E. Stark
Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109-2122
{worthena, stark}@eecs.umich.edu

Abstract

We develop a belief-propagation-based joint decoding and channel state estimation algorithm for low-density parity check codes on fading channels with memory. The excellent empirical performance of LDPC codes on the AWGN channel and the possibility of arranging the code memory relative to the channel correlation suggests that LDPC codes have the potential to perform well on fading channels with memory. We treat the block fading channel, which models a frequency-hopped spread spectrum system. We present performance results for simple codes decoded using our algorithm and compare them with the performance of the belief propagation algorithm without channel state estimation and block-interleaved Reed-Solomon codes.

1 Introduction

Low-density parity check (LDPC) codes, first proposed by Gallager [1, 2], have recently been found to exhibit near Shannon capacity performance on the additive white Gaussian noise (AWGN) channel using a practical decoder [3]. Some LDPC codes over $GF(q)$ with $q > 2$ have been shown to have performance rivaling that of the best known turbo codes [4, 5]. The codes are specified by very sparse random parity check matrices in which each coded symbol participates in no more than j parity check equations and each parity check equation uses no more than $k > j$ coded symbols where j and k are small integers irrespective of the block length n. Uniform column and row weights have typically been used, but probably do not give the best performance. The code is also compactly described by a bipartite graph in which each row or column becomes a vertex, and a 1 in the parity check matrix indicates that the corresponding row and column vertices are connected by an edge.

Numerous algorithms have been proposed for generating random matrices (or equivalently graphs) with desired weights and minimal column overlap. Sipser and Spielman [6] give a construction of a more general set of "expander codes" based on expander graphs

---

*This research was supported by the Army Research Office under Grant DAAH04-96-1-0110 and by a National Science Foundation Graduate Research Fellowship.

whose expansion property seems to impart good properties to the resulting codes. They also describe algorithms for generating random graphs likely to have good expansion.

The codes can be efficiently decoded using an iterative belief propagation algorithm which approximates maximum likelihood decoding. This algorithm, which McEliece, MacKay, and Cheng [7] have shown to be related to the well-known turbo decoding algorithm, successively updates estimates of the *a posteriori* symbol probabilities using the graph associated with the parity check matrix. Although no convergence proof for this algorithm exists so far in the case of graphs with "loops", numerous empirical results for turbo codes and LDPC codes show that it performs very well. Previous work on belief propagation appears in the artificial intelligence literature on Bayesian networks [8].

The excellent performance of LDPC codes on the AWGN channel and the proven performance of turbo codes with iterative decoding motivates the extension of these ideas to fading channels with memory. Because LDPC codes are based on random bipartite graphs, carefully chosen constraints on the graph construction could yield codes especially suited to particular fading channel applications. For example, block interleaving can be built directly into the code. Also, when the fading level (or time-varying interference power) is unknown, joint decoding and channel state estimation is optimal. We show that the LDPC decoding algorithm is easily extended to include channel state estimation and characterize its performance on a simplified fading channel model for frequency-hopped spread spectrum systems.

In section 2, we describe the system model. Section 3 describes the decoding algorithm for fading channels. Section 4 gives the results of some performance simulations. In section 5, we present our conclusions.

2 System Model

We investigate iterative decoders for LDPC codes transmitted over a simplified, correlated fading channel model. We consider the block fading channel which models frequency-hopped spread spectrum (FH-SS) systems. We model this channel as a pair of symmetric Gaussian channels with different signal-to-noise ratios (SNR's) employing binary antipodal signaling with coherent detection. For each block of m coded symbols, we choose one channel with probability p and the other with probability $1-p$ and transmit the symbols over that channel. If the receiver knows which channel was used for each symbol, we say that perfect channel state information (CSI) is available. Otherwise, we assume that the receiver can deduce the channel state only from the received symbols. In both cases, we assume that the receiver knows the block boundaries, channel selection probabilities, and channel SNR's exactly. In a FH-SS system, the "good" and "bad" channels correspond to hops enhanced by constructive multi-path interference or degraded by destructive multipath (or jammers or multiple access interference) respectively, and the pseudo-random hopping pattern ensures that the block states are independent.

3 Decoding Algorithm

Consider the symbol-by-symbol maximum *a posteriori* probability (MAP) decision rule for a symbol v_i given some statistically related evidence $\underline{x}$

$$\hat{v}_i = \arg\max_{v_i} \mathrm{p}(v_i|\underline{x}).$$

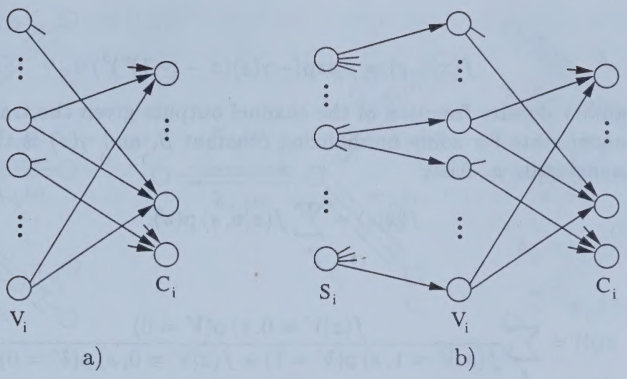


Figure 1: Belief networks for LDPC codes. a) network for AWGN channel. In this example, each variable is the parent of two constraints and each constraint is the child of three variables. b) network for block interference channel. Here each channel state is the parent of four consecutive variables and each variable is the child of exactly one channel state.

Previous literature has shown that Pearl's belief propagation algorithm [8] can be used to approximately solve this problem for LDPC codes. We extend the application of Pearl's algorithm to joint decoding and channel state estimation as follows.

Figure 1a shows the belief network for an LDPC code on an AWGN channel[1]. Following the nomenclature of Sipser and Spielman [6], "variables" V correspond to unknown coded symbols and "constraints" C correspond to parity check equations. We say that a constraint is satisfied if the sum mod 2 of its parent variables' values is 0. In Figure 1b, we extend this belief network to include channel state estimate nodes S for the block fading channel. In both cases, the observed channel outputs are incorporated into the variable nodes and are either conditionally independent given the transmitted symbol for the AWGN channel, or conditionally independent given the channel state and the transmitted symbol for the block fading channel. Each node Z holds two vectors $\pi(z)$ and $\lambda(z)$ corresponding to likelihood contributions for each of its possible values from its parents and children respectively. The algorithm updates these values by combining messages $\pi_Z(u_i)$ from the node's parents and $\lambda_{Y_i}(z)$ from the node's children. The total likelihood estimate is $\mathrm{p}(z|\underline{x}) = \alpha\pi(z)\lambda(z)$ for some suitable normalizing constant α. In practice, the nodes are updated repeatedly until a "good" estimate of the MAP decision is reached.

The channel node values $\pi(s)$ and $\pi_{V_i}(s)$ are initialized to the *a priori* probabilities of the channel states p and $1-p$. The variable node values $\pi(v)$ and $\pi_{C_i}(v)$ are initialized to the *a posteriori* probabilities of the coded symbols given the observed channel outputs x and the priors on the channel states $\mathrm{p}(s)$:

$$\pi(v) = \mathrm{p}(v|x) = \sum_{s} \frac{f(x|v,s)\,\mathrm{p}(v|s)}{f(x|s)}\,\mathrm{p}(s)$$

---

[1] This is analogous to Figure 10 of [7].

where

$$f(x|v,s) = \beta \exp(-\gamma(s)(x-(-1)^v)^2)$$

is the probability density function of the channel outputs given the transmitted symbol and the channel state for some normalizing constant β, and $\gamma(s)$ is the coded symbol SNR for channel state s. Also,

$$f(x|s) = \sum_v f(x|v,s)\,\mathrm{p}(v).$$

This gives

$$\pi(0) = \sum_s \frac{f(x|V=0,s)\,\mathrm{p}(V=0)}{f(x|V=1,s)\,\mathrm{p}(V=1) + f(x|V=0,s)\,\mathrm{p}(V=0)}\,\mathrm{p}(s)$$

$$= \sum_s \frac{1}{e^{-4x\gamma(s)}+1}\,\mathrm{p}(s)$$

$$\pi(1) = 1-\pi(0) = \sum_s \frac{e^{-4x\gamma(s)}}{e^{-4x\gamma(s)}+1}\,\mathrm{p}(s)$$

where $\mathrm{p}(v)$, the prior on the sent symbol value $v \in \{0,1\}$, was taken to be 1/2. Note that the channel values x are normalized so that $\mathrm{E}[x|{+1}\text{ sent}] = 1$. We update the variables before the channel states so it is unnecessary to initialize $\lambda_V(s)$. The constraints are known to be satisfied so we just initialize the $\lambda_C(v_i)$ to be equally likely (i.e. the constraints do not contribute information about the variable states without information about other variables).

For the channel state update, we calculate

$$\lambda(s) = \prod_i \lambda_{V_i}(s) \qquad \text{for each possible } s$$

and the values needed by the variables, which are updated priors on the channel state

$$\pi_{V_i}(s) = \pi(s) \prod_{k \neq i} \lambda_{V_k}(s)$$

(see Figure 2a). Note that $\lambda_{V_i}(s)$ is not included since this would be reusing information and result in undesired positive feedback.

For the variable update, we calculate

$$\lambda(v) = \prod_i \lambda_{C_i}(v) \qquad \text{for each possible } v$$

and

$$\pi(v) = \sum_s \mathrm{p}(v|x,s)\pi_{V_i}(s)$$

so

$$\pi(0) = \sum_s \frac{1}{e^{-4x\gamma(s)}+1}\pi_{V_i}(s)$$

$$\pi(1) = 1-\pi(0) = \sum_s \frac{e^{-4x\gamma(s)}}{e^{-4x\gamma(s)}+1}\pi_{V_i}(s).$$

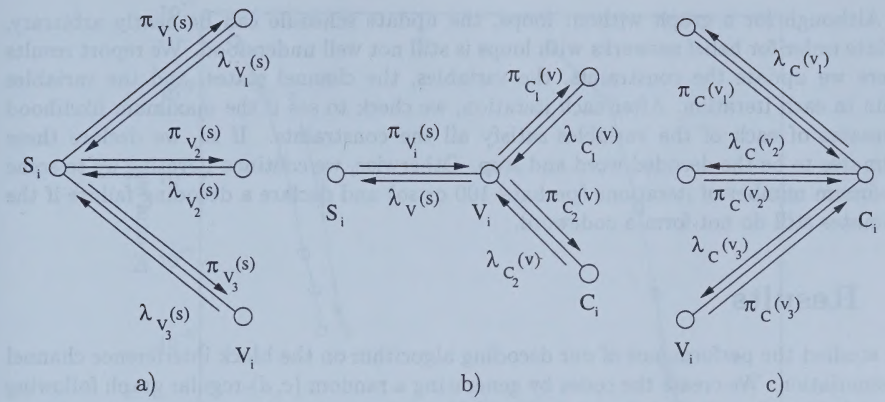


Figure 2: Messages for node updates: a) channel state node, b) variable node, c) constraint node

The messages to be passed to the variable's neighbors are

$$\pi_{C_i}(v) = \pi(v) \prod_{k \neq i} \lambda_{C_k}(v),$$

which are updated estimates of $\mathrm{p}(v|x, s, \text{other constraints})$, and the updated state estimate contributions

$$\begin{aligned} \lambda_V(s) &= \sum_v \lambda(v)\,\mathrm{p}(v|s) \\ &= \lambda(0)\frac{1}{e^{-4x\gamma(s)}+1} + \lambda(1)\frac{e^{-4x\gamma(s)}}{e^{-4x\gamma(s)}+1} \end{aligned}$$

(see Figure 2b).

Finally, for the constraint update, we need calculate only the values $\lambda_C(v_i)$ which will be needed in the next variable update. The local values are irrelevant since we know that the constraint must be satisfied. The $\lambda_C(v_i)$ represent the evidence that other variables participating in constraint C contribute to our knowledge of the variable V_i. Thus, from Pearl's formulas,

$$\lambda_C(v_i) = \sum_c \lambda(c) \sum_{\underline{v}:[\underline{v}]_i = v_i} \mathrm{p}(c|\underline{v}) \prod_{k \neq i} \pi_C(v_k).$$

However, to be consistent with the constraint being satisfied $\lambda(c)$ must equal 1 only when $c = 0$. Similarly, we can interpret $\mathrm{p}(C = 0|\underline{v})$ as being 1 if the parity of $\underline{v}$ is even and 0 otherwise, so

$$\lambda_C(v_i) = \sum_{\underline{v}:[\underline{v}]_i = v_i} I\{\underline{v} \text{ has even parity}\} \prod_{k \neq i} \pi_C(v_k).$$

$\underline{v}$ represents the vector of states of C's parents, $I\{\cdot\}$ is the indicator function, and the sum is over all possible states such that the ith state equals v_i (see Figure 2c).

Although for a graph without loops, the update schedule can be nearly arbitrary, update order for belief networks with loops is still not well understood. We report results where we update the constraints, the variables, the channel states, and the variables again in each iteration. After each iteration, we check to see if the maximum likelihood estimates of each of the variables satisfy all the constraints. If so, we declare these estimates to be the decoded word and stop. Otherwise, we continue iterating up to some maximum number of iterations (perhaps 100 or so) and declare a decoding failure if the estimates still do not form a codeword.

4 Results

We studied the performance of our decoding algorithm on the block interference channel by simulation. We create the codes by generating a random (c, d)-regular graph following the technique described in Appendix II of [6] and eliminating the multi-edges by random swaps. Graphs created this way are likely to have good expansion and thus may give good performance. This graph becomes the variable-to-constraint graph in the decoding algorithm. We encode a selection of equally-likely random bits using a possibly permuted, systematic generator matrix, pass the resulting codeword through a simulated channel, decode the received channel outputs using our decoding algorithm, and report the number of information bit errors. All the information bit errors we observed were caused by decoding failures.

Figure 3 shows our simulation results for a block length 10,240 code with 5,120 information bits (and thus rate 1/2) on a block fading channel. The code was constructed as described above such that each variable participates in 3 constraints and each constraint depends on 6 variables. No attempt was made to optimize the construction. We compare the results for our algorithm with an algorithm that uses only the prior probabilities of the channel states ("no channel estimation") and with a Reed-Solomon code with compatible parameters. The gain of channel estimation over no channel estimation is roughly 0.2 dB. This improvement might be more with other channel parameters. The gain of the LDPC code over the RS code, which has the same block length and rate and is very robust to bursts of errors in a single hop, is about 3 dB. About 2 dB of this gain is because LDPC codes use soft-inputs; the rest is due to improved code performance. Further gains may be possible by using more carefully designed codes or making algorithm modifications.

Figure 4 shows a histogram of the number of iterations required to decode a codeword. This shows that despite the fact that the average number of iterations per word was only 20.9, there were a significant fraction of codewords that required more than 50 iterations to decode. Note that the right-most bin includes the words that ran to 100 iterations and failed to decode. These results underscore the importance of running the algorithm until a codeword is found even if it takes 100 or more iterations. This makes real-time implementation more difficult so approaches to either reducing the fraction of words that require a large number of iterations to decode or accelerating the algorithm would be useful.

In Figure 5, we compare the performance of LDPC codes with channel estimation with the SINCGARS RS coding system. The SINCGARS code block-interleaves 24 (32,12) extended RS codewords and adds a parity check to each symbol to form a single block of 4608 coded bits with a hop length of 144 bits (one symbol plus its parity check from each codeword). The decoder uses the parity checks to generate erasures for errors and

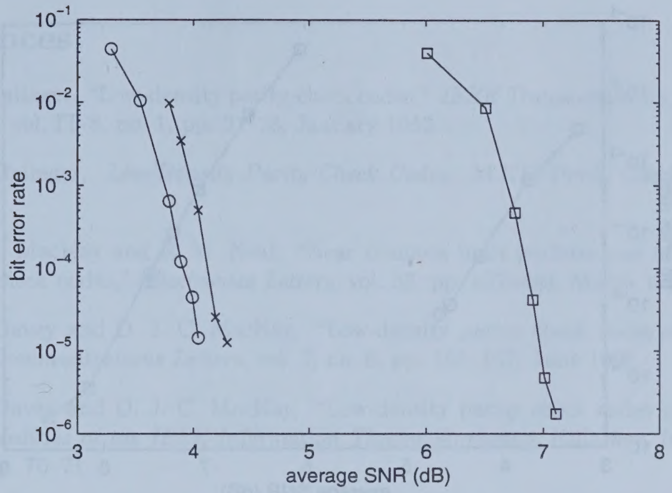


Figure 3: Performance of (10240,5120) LDPC code on the block interference channel. ○ LDPC code with channel estimation, × LDPC code without channel estimation, □ (1024,512) extended RS code. Channel SNR's were separated by 7 dB, the probability of choosing the good channel was 0.5, and the hop length was 10 bits.

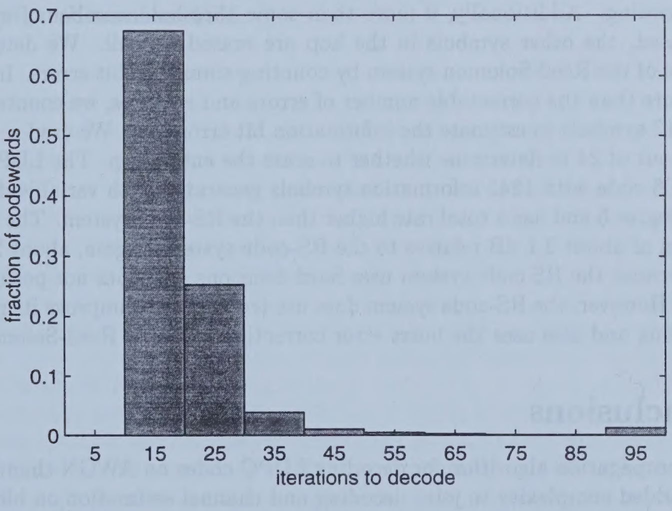


Figure 4: Iteration histogram for (10240,5120) LDPC code on the block interference channel. ○ LDPC code with channel estimation. Channel SNR's were separated by 7 dB, the average SNR was 3.8 dB, the probability of choosing the good channel was 0.5, and the hop length was 10 bits.

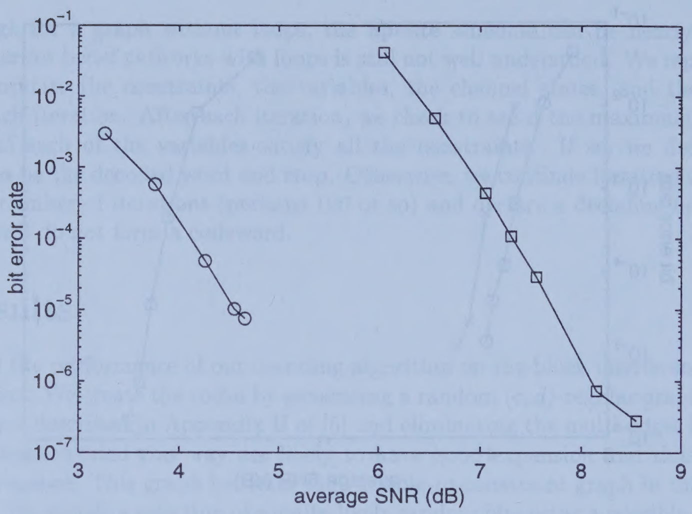


Figure 5: Performance of (4608,1843) LDPC code on the block interference channel. ○ LDPC code with channel estimation, ☐ SINCGARS RS code. Channel SNR's were separated by 7 dB, the probability of choosing the good channel was 0.5, and the hop length was 144 bits.

erasures decoding. Additionally, if more than some threshold number of symbols in a hop are erased, the other symbols in the hop are erased as well. We determined the performance of the Reed-Solomon system by counting simulated bit errors. In codewords that had more than the correctable number of errors and erasures, we counted bit errors in the first 12 symbols to estimate the information bit error rate. We used a threshold of 20 symbols out of 24 to determine whether to erase the entire hop. The LDPC code was a length 4608 code with 1843 information symbols generated with variable degree 3 and constraint degree 5 and has a total rate higher than the RS-code system. The LDPC code shows a gain of about 3.1 dB relative to the RS-code system. Again, about 2 dB of this might be because the RS code system uses hard decisions and does not perform channel estimation. However, the RS-code system does use techniques to improve its performance in block fading and also uses the burst error correction power of Reed-Solomon codes.

5 Conclusions

The belief propagation algorithm for decoding LDPC codes on AWGN channels extends with little added complexity to joint decoding and channel estimation on block interference channels. We have shown that the joint decoding and channel estimation algorithm performs somewhat better than decoding using only the channel state *a priori* probabilities. Also, LDPC codes perform better than comparable Reed-Solomon code-based systems even though the Reed-Solomon codes have good burst error correction capability.

References

[1] R. G. Gallager, "Low-density parity-check codes," *IEEE Transactions on Information Theory*, vol. IT-8, no. 1, pp. 21–28, January 1962.

[2] R. G. Gallager, *Low-Density Parity-Check Codes*, M.I.T. Press, Cambridge, MA, 1963.

[3] D. J. C. MacKay and R. M. Neal, "Near shannon limit performance of low density parity check codes," *Electronics Letters*, vol. 33, pp. 457–458, March 1997.

[4] M. C. Davey and D. J. C. MacKay, "Low-density parity check codes over GF(q)," *IEEE Communications Letters*, vol. 2, no. 6, pp. 165–167, June 1998.

[5] M. C. Davey and D. J. C. MacKay, "Low-density parity check codes over GF(q)," in *Proceedings of the IEEE Information Theory Workshop*, Killarney, Ireland, June 1998, pp. 70–71.

[6] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1710–1722, Nov. 1996.

[7] R. J. McEliece, D. J. C. MacKay, and J-F. Cheng, "Turbo decoding as an instance of Pearl's "belief propagation" algorithm," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 2, pp. 140–152, February 1998.

[8] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

Coded Modulation for Peak Power Constrained Wireless OFDM Systems[1]

Dennis L. Goeckel
University of Massachusetts - Amherst
ECE Department
Box 35110
Amherst, MA 01003-5110
e-mail: goeckel@ecs.umass.edu

Abstract

This paper considers the design of coded modulation schemes that provide both error correction and control of the envelope of the transmitted signal in wireless orthogonal frequency division multiplexing (OFDM) systems. The standard OFDM restriction of employing identical signal sets on each subcarrier is removed, and a construction method for the resulting frequency domain coded modulation (FDCM) system is described. A sample construction exhibits a 5.4 to 8.5 dB gain in required peak signal-to-noise ratio over a standard OFDM system of similar complexity.

1 Introduction

A key inhibiting factor to the realization of high-speed ubiquitous access to information is the wireless link, which is plagued by limited bandwidth availability and the presence of multipath fading. Two highly regarded architectures for high-speed (wired or wireless) data communications are: (1) a coded single-carrier system employing a decision feedback equalizer (DFE) at the receiver, and (2) a coded orthogonal frequency division multiplexing (OFDM) system. Coded OFDM systems have been a topic of significant interest recently, finding application in digital audio broadcast (DAB) and digital video standards (DVB) in Europe, and in asynchronous digital subscriber lines (ADSL) in the United States. However, the applicability of OFDM for low-power two-way wireless applications has been questioned by factors that were not critical in the design of the DAB, DVB, nor ADSL systems; in particular, to avoid the deleterious effects of in-band signal distortion and out-of-band spectral regrowth, efficient amplifier operation precludes baseband signals such as the OFDM signal that possess a large peak-to-mean envelope power ratio (PMEPR) [1]. Partly because of the significance of the PMEPR problem, OFDM has been bypassed for high-speed wireless local area network (WLAN) standards [2], and there has been significant work to remedy this shortcoming of OFDM in recent years (e.g. [2, 3, 4, 5, 6]).

One possible solution to the PMEPR problem is to employ a block coding scheme $C : \{0,1\}^k \rightarrow \{0,1\}^N$ that restricts the input of the N-point inverse fast Fourier transform

---

[1] This paper is based in part upon work supported by the National Science Foundation under Grant No. NCR-9714597.

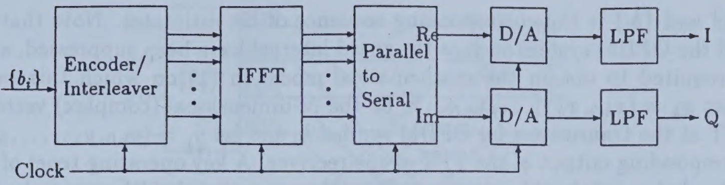


Figure 1: Baseband portion of the transmitter for a FDCM system.

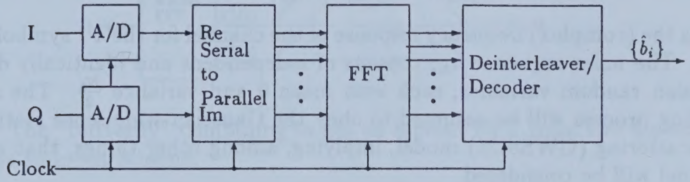


Figure 2: Baseband portion of the receiver for a FDCM system.

(IFFT) at the OFDM transmitter to contain only sequences that do not produce large peaks in the envelope of the transmitted signal [2, 3, 4, 5]. In these methods, researchers have followed the basic OFDM tenet of employing a standard single-carrier signal set (e.g. binary phase-shift keying (BPSK)) on each of the subcarriers. We argue that this basic OFDM assumption - that identical signal sets typical of single-carrier systems must be employed on each subcarrier - can be removed if complexity is carefully monitored; in other words, the restriction of the input of the IFFT in the OFDM transmitter to the subspace $\{0,1\}^N$ of C^N, where C is the space of complex numbers, is unnecessary. This restriction is carried over from a single-carrier system; however, in an OFDM system, restricting signal sets to lie in $\{0,1\}^N$ does not generally aid in synchronization and does not directly aid in maintaining a constant envelope of the signal. Thus, signal points will be drawn arbitrarily from the signal space, which can be predicted to allow a significant increase in the diversity achieved by the system, as was observed by previous authors who have used similar constructions for a single-carrier system (e.g. [7, 8]). However, in contrast to the work in [7, 8], such an approach will allow considerable freedom in controlling the envelope of the transmitted OFDM signal, which will lead to a reduction in the linearity requirements of the amplifier. Since identical signal sets will not be employed on each of the subcarriers, the new system is denoted a frequency domain coded modulation (FDCM) system.

In Section 2, the system assumptions and optimization criterion are described. Section 3 considers the general construction of FDCM, and Section 4 presents a sample construction with the associated numerical results and analysis. Finally, Section 5 presents the conclusions and continuing work.

2 Definition of the Problem

2.1 System Assumptions

The major baseband elements of the transmitter and receiver for a frequency domain coded modulation (or OFDM) system employing N subcarriers are shown in Figures 1 and 2, respectively, where $\{b_i\}$ is the sequence of information bits to be transmitted across the channel and $\{\hat{b}_i\}$ is the corresponding sequence of bit estimates. Note that certain elements of the OFDM system such as the guard interval have been suppressed, although these are required to obtain the mathematical model in (1) on which this work will operate. Let $\underline{x}_k = (x_{k,0}, x_{k,1}, \ldots, x_{k,N-1})^T$ be the N-dimensional (complex) vector input to the IFFT at the transmitter for OFDM symbol k, and let $\underline{y}_k = (y_{k,0}, y_{k,1}, \ldots, y_{k,N-1})^T$ be the corresponding output of the FFT at the receiver. A key operating tenet of OFDM is that each subchannel should occupy a sufficiently narrow bandwidth so as to be affected by frequency nonselective fading; thus,

$$y_{k,i} = H_{k,i} x_{k,i} + n_{k,i}, \tag{1}$$

where $H_{k,i}$ is the (complex) frequency response of the channel for the k^{th} symbol on the i^{th} subchannel. The noise sequence $n_{k,i}$ consists of independent and identically distributed (IID) Gaussian random variables, each with mean 0 and variance $\frac{N_0}{2}$. The zero-mean channel fading process will be assumed to obey the Gaussian wide-sense stationary uncorrelated scattering (GWSSUS) model, implying, among other things, that a Rayleigh fading channel will be considered.

2.2 Optimization Criterion

Assuming maximum likelihood sequence detection with perfect channel measurements at the receiver, the goal is to find the mapping from the input bit sequence $\{b_i\}$ to the sequence $\{\underline{x}_k\}$ that maximizes performance with reasonable decoding complexity. To derive the performance measure of interest, let $\zeta(C)$, the PMEPR of a coded modulation strategy C, be defined as follows. If ξ is the set of all possible vectors $\underline{x}_k$ under C, then,

$$\zeta(C) = \frac{\max_{\underline{x}_k \in \xi} \max_j |X_N^k(j)|^2}{E_{\underline{x}_k}\left[\frac{1}{N}\sum_{j=0}^{N-1} |X_N^k(j)|^2\right]}, \tag{2}$$

where $X_N^k(j)$ is the j^{th} element of the N-point FFT of the vector $\underline{x}_k$. Strategies are compared based on the peak signal-to-noise ratio (SNR) required to achieve the bit error probability P_b of interest. Denoting $\frac{E_s}{N_0}(C)$ as the average signal-to-noise ratio required by strategy C to achieve P_b, the performance metric is given by:

$$\frac{E_p}{N_0}(C) = \frac{E_s}{N_0}(C) + \zeta(C), \tag{3}$$

where all quantities are in dB. This is an overly simple optimization criterion, as it implies that the amount of backoff of the power amplifier is proportional to the PMEPR. However, it yields a simple and reasonable criterion by which various coded modulation strategies can be compared.

3 FDCM System Characterization

Consider a set $S = \{\underline{s}_i : i = 0, \ldots, K-1\}$ of signal vectors in M complex dimensions (i.e. $\underline{s}_i = (s_{i,0}, s_{i,1}, \ldots, s_{i,M-1})^T$) that has been designed such that, among other attributes, the largest magnitude of an element of the M-point FFT[2] of the vector $\underline{s}_i$ is "small" for all i; that is, $\max_i \max_n |\sum_{k=0}^{M-1} s_{i,k} e^{-j2\pi\frac{kn}{M}}|^2$ is less than some reasonable constant. Before considering the choice of S, consider how signal vectors drawn from this signal set can be input to an N-point FFT at the transmitter in a FDCM system so as to: (1) control the PMEPR, and (2) maximize the diversity achieved by the system

---

[2] Although it is strictly the IFFT in which we are interested, its structure is, of course, virtually identical to the FFT and thus admits the same recipe.

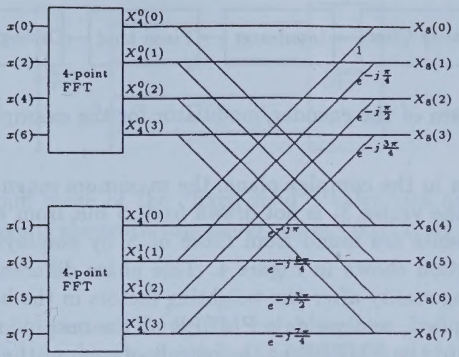

x(0)
x(2)
x(4)
x(6)
4-point FFT
$X_4^0(0)$
$X_4^0(1)$
$X_4^0(2)$
$X_4^0(3)$
x(1)
x(3)
x(5)
x(7)
4-point FFT
$X_4^1(0)$
$X_4^1(1)$
$X_4^1(2)$
$X_4^1(3)$
1
$e^{-j\frac{\pi}{4}}$
$e^{-j\frac{\pi}{2}}$
$e^{-j\frac{3\pi}{4}}$
$e^{-j\pi}$
$e^{-j\frac{5\pi}{4}}$
$e^{-j\frac{3\pi}{2}}$
$e^{-j\frac{7\pi}{4}}$
$X_8(0)$
$X_8(1)$
$X_8(2)$
$X_8(3)$
$X_8(4)$
$X_8(5)$
$X_8(6)$
$X_8(7)$

Figure 3: The "butterfly" combining to get an 8-point FFT from two 4-point FFTs in a decimation-in-input scheme.

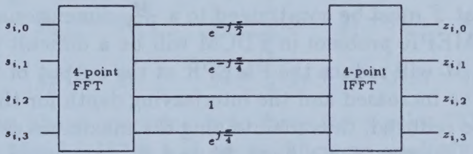

$s_{i,0}$
$s_{i,1}$
$s_{i,2}$
$s_{i,3}$
4-point FFT
$e^{-j\frac{\pi}{2}}$
$e^{-j\frac{\pi}{4}}$
1
$e^{j\frac{\pi}{4}}$
4-point IFFT
$z_{i,0}$
$z_{i,1}$
$z_{i,2}$
$z_{i,3}$

Figure 4: Method for obtaining element $\underline{z}_i$ of the phase modified version of S from element $\underline{s}_i$ of S.

by separating subcarriers carrying the dimensions of the signal vector as far apart as possible in frequency.

For exposition purposes, assume that $M = 4$ and that $N \geq 8$ is equal to a power of two. To find the full N-point FFT, a decimation-in-time algorithm [9, pg. 291] requires as one stage the formation of 8-point FFTs. Figure 3 shows the final step for one such formation, which combines the outputs of two 4-point FFTs, denoted $\{X_4^0(k)\}$ and $\{X_4^1(k)\}$, respectively, to form the 8-point FFT $\{X_8(k)\}$. Suppose the 4-dimensional signal vectors $\underline{\tilde{s}}_0$ and $\underline{\tilde{s}}_1$, which are selected from S based on the information bits, are placed across the input of the upper and lower 4-point FFTs shown in Figure 3, respectively. Note that this implies an interleaving by two subcarriers of the dimensions of each of the signal vectors. Consider the outputs $X_8(0)$ and $X_8(4)$ of the 8-point FFT, which are given by:

$$\begin{aligned} X_8(0) &= X_4^0(0) + X_4^1(0) \\ X_8(4) &= X_4^0(0) + e^{-j\pi} X_4^1(0) \end{aligned} \tag{4}$$

Clearly $|X_8(0)| \leq |X_4^0(0)| + |X_4^1(0)|$ and $|X_8(4)| \leq |X_4^0(0)| + |X_4^1(0)|$, which implies that controlling the maximum magnitude of an element of the 4-point FFTs does control the maximum magnitude of an element in the 8-point FFT; however, we can do much better. A simple consideration of (4) reveals that the maximum of the magnitudes of $X_8(0)$ and $X_8(4)$ is minimized (for $X_4^0(0)$ and $X_4^1(0)$ of fixed magnitudes) when $X_4^0(0)$ and $X_4^1(0)$ are orthogonal in the complex plane; that is, $\mathrm{Re}\{X_4^0(0)X_4^1(0)^*\} = 0$, where $\mathrm{Re}(z)$ denotes the real part of z and $X_4^1(0)^*$ is the complex conjugate of $X_4^1(0)$. Generalizing this to all $X_8(k)$, if each of $X_4^0(k)$ and $X_4^1(k)$, $\forall k$, are constrained (across all possible input bits)

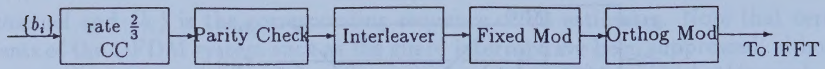


Figure 5: Block diagram of the encoder/modulator for the example of Section 4.

to only a single dimension in the complex plane, the maximum magnitude of $X_8(k)$ can be minimized for all k if the vector $\tilde{\underline{s}}_1$ is not drawn from S but from a "phase modified" version of S, whose elements are found from those of S by employing the Euclidean-distance preserving operation shown in Figure 4. This phase difference between $\tilde{\underline{s}}_0$ and $\tilde{\underline{s}}_1$ forces the desired orthogonality after the weighting factors in the butterfly combining of Figure 3 have been applied, and yields a PMEPR at the output of the 8-point FFT that is less than or equal to the PMEPR at the output of each of the 4-point FFTs.

The above construction can be applied for arbitrary M to guarantee that the PMEPR at the output of a $2M$-point FFT is less than or equal to that at the output of its two M-point constituent FFTs. Next briefly consider controlling the PMEPR at the output of a $2^l M$-point FFT, $l > 1$. If analogous methods to those discussed above are employed, limiting the PMEPR at the output of the $2^l M$-point FFT to that at the output of the M-point FFT requires that S must be constrained to a $\frac{M}{2^{l-1}}$-dimensional subspace of $\mathcal{C}^M$. This suggests that the PMEPR problem in FDCM will be a difficult problem to solve. For a fixed N, increasing M will reduce the PMEPR at the output of the N-point FFT but both complexity will be increased and the interleaving depth for the dimensions of a single signal vector will be reduced, thus maintaining the maximum diversity achievable without further coding. Likewise, for a fixed M and N, increasing l will reduce the PMEPR at the output of the N-point FFT but the reduced dimensionality of the signal set S will necessitate a decrease in the minimum Euclidean distance of two signal vectors of S for a fixed average signal-to-noise ratio. However, the following recipe is suggested:

1. Find a signal set $S = \{\underline{s}_i : i = 0, \ldots, K-1\}$ of signal vectors in M complex dimensions such that both the signal set S and its phase modified version(s), if any, exhibit a large minimum product distance between any two vectors. Symbols will be drawn alternately from S and its phase modified version(s).
2. Block interleave the dimensions of the signal vectors to a depth of N/M, thus implying that the M dimensions of a signal vector will go into the same M-point FFT, which is combined with the M-point FFTs of other symbols through butterflies to get the full N-point FFT.
3. Design an outer code with output alphabet S. Interleave the outer code symbols in time to provide time diversity.

4 An FDCM Example

4.1 Construction

The construction shown in Figure 5 is based on the development of Section 3 and two observations: (1) forming S from a product of binary signal sets will lead to computational savings, (2) the work in [2] indicates that forming signal vectors $\underline{s}_i$ by choosing signals of odd parity in $\{0,1\}^4$, followed by the mapping $0 \rightarrow +1$ and $1 \rightarrow -1$, will result in a small PMEPR at the output of the 4-point FFT of the signal vectors.

The system works as follows. Information bits are input to a rate $\frac{2}{3}$, constraint length 5, convolutional code. Next, a redundant bit is added to each set of three output bits to

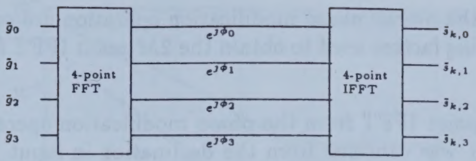


Figure 6: The second stage of the "fixed mod": calculates an element of S from the vector $\tilde{\underline{g}}$. For the results shown in this paper, $\phi_1 = 4.69772$, $\phi_2 = 5.89327$, $\phi_3 = 5.50689$, and $\phi_4 = 5.09938$.

form a four-bit symbol vector. Using this redundant bit, the odd parity of each of the four-bit symbols can be guaranteed, while the free distance of the code is increased from 5 to 6. The resulting four-bit symbols are then interleaved (at the symbol level) at an arbitrary depth. The mapping 0 → +1 and 1 → −1 is then performed, which results in a four-dimensional symbol vector $\underline{g} = (g_0, g_1, g_2, g_3)^T$. Finally, $\underline{g}$ is transformed into a signal vector from S by applying the "fixed mod", which consists of two steps. The first step converts $\underline{g}$ to $\tilde{\underline{g}}_i$ as:

$$\begin{aligned} \tilde{g}_0 &= g_0 \qquad \tilde{g}_1 = \frac{1}{2}(g_1 + g_3 + j(g_1 - g_3)) \\ \tilde{g}_2 &= g_2 \qquad \tilde{g}_3 = \frac{1}{2}(g_1 + g_3 + j(g_3 - g_1)). \end{aligned}$$

The purpose of the above operation is to force the 4-point IFFT of $\tilde{g}$ to be real. The second step of the "fixed mod" is a phase modification applied to $\tilde{\underline{g}}$ as shown in Figure 6. This phase modification, which was found by computer search, is chosen to maximize the minimum product distance between two possible symbols at the output of the orthogonality phase modifier. Finally, the orthogonality phase modifier applies the operation shown in Figure 4 (with phase multipliers modified to account for the difference between the FFT and IFFT) to 4-dimensional symbols of odd index (i.e. every other symbol). The purpose of the orthogonality phase modifier is to force the desired orthogonality in the combining that forms the 8-point FFTs per Section 3. From the output of the coder in Figure 5, $\frac{N}{4}$ symbols are taken at at time, interleaved at depth $\frac{N}{4}$, and are used to form $\underline{x}_k$, the input to the N-point IFFT at the transmitter of the OFDM system.

The constructed code will yield a PMEPR of 0 dB at the output of the 8-point FFTs, which will increase to $3\log_2 \frac{N}{8}$ dB for a system with N subcarriers. The dimensions of a signal from S are interleaved at a depth of $2\frac{N}{8}$ subcarriers, whereas the 4-bit symbols of the outer code can of course be interleaved at an arbitrary depth without affecting the PMEPR.

4.2 Complexity

Regardless of the value of M, the second stage of the "fixed mod" and the "orthogonality mod" can be implemented together with three steps: (1) M-point IFFT, (2) multiplication by appropriate phase factors, and (3) M-point FFT. To consider the impact on system complexity, first consider the transmitter, where the structure of a decimation-in-input N-point IFFT will begin with $\frac{N}{M}$ M-point IFFTs (analogously to the decomposition of the FFT shown in Figure 3). The following steps result in a simplified structure:

1. The M-point FFT at the conclusion of the phase modification cancels with the M-point IFFT that is the first step of the decimation-in-input N-point IFFT.

2. The multipliers of the overall phase modification operation are combined with the "butterfly" combining factors used to obtain the $2M$ point IFFT from two M point IFFTs.

3. The remaining M-point IFFT from the phase modification operation replaces the M-point IFFT that was canceled from the decimation-in-input N-point IFFT in step 1.

The net effect of these three steps, which combine both phase modification operations with the decimation-in-time N-point IFFT, is a modification of the combining factors to form the $2M$ point IFFTs; thus, there is no complexity increase at the transmitter due to the phase modification of the symbols. Next consider the receiver. It can be shown that applying the phase modification operations does not alter the Euclidean distance between two signals. Thus, if no subchannel fading information (i.e. channel side information (CSI)) is available, the FFT output is processed by undoing the phase modification operation and then decode using the standard Viterbi algorithm with a simple add-compare-select update at each state (note that this was the main reason for basing the algorithm around the binary convolutional code). Similarly to the transmitter, the undoing of the phase modification operations can be worked into the N-point FFT by using a decimation-in-output [9, pg. 302] structure for the N-point FFT at the receiver; thus, if no subchannel fading information is employed, there is no complexity increase at the receiver. However, when subchannel fading information is used at the receiver, the straightforward approach to branch metric calculation leads to $16N$ extra scalar multiplications per OFDM symbol, which is equivalent to $4N$ extra complex multiplications per OFDM symbol, that must be performed at the receiver. For a reasonably large N, this does not greatly increase complexity over that required for the FFT.

4.3 Numerical Results

Simulations were performed for three cases. The first case assumes independent fading between subchannels dimensions of a signal from S (i.e. assumes a separation of $2\frac{N}{8}$ subchannels is far enough to provide independent fading). For the second and third cases, simulations were performed using the 6-tap GSM model parameters for an urban area with $N = 16$ and $N = 64$, respectively. Perfect interleaving is assumed for the 4-bit symbols from the encoder (but, of course, not for the individual dimensions of the 4-dimensional symbols). The decoding complexity of the scheme is calculated as follows: the total number of branches leaving all states times the number of complex symbols per branch is divided by the number of information bits per trellis transition. This yields a complexity of $\frac{2^4 \times 4 \times 4}{2} = 128$ for the proposed scheme. Simulation results are shown in Figures 7, 8, and 9 for the proposed scheme along with those for a perfectly interleaved (at the bit level) rate $\frac{1}{2}$ constraint length 6 convolutional code whose output is mapped onto a BPSK signal set, which will have similar complexity to the proposed scheme by the measure proffered above. Note that as N increases, the PMEPR of the proposed system increases, but the PMEPR gain over the standard scheme remains the same. At $P_b = 10^{-4}$, the proposed scheme requires approximately 0.5 to 1.5 dB more *average* signal-to-noise ratio than the standard system. However, the proposed scheme shows a PMEPR gain of 9.0 dB over the standard scheme, indicating a gain in required peak signal-to-noise ratio of 7.5 to 8.5 dB using this simple construction. Next, recall from Section 4.2 that to employ channel state information, the proposed scheme currently requires $4N$ extra complex multiplications per OFDM symbol. Suppose that this additional complexity is undesirable; in this case, the proposed scheme can run with no CSI, in which case it

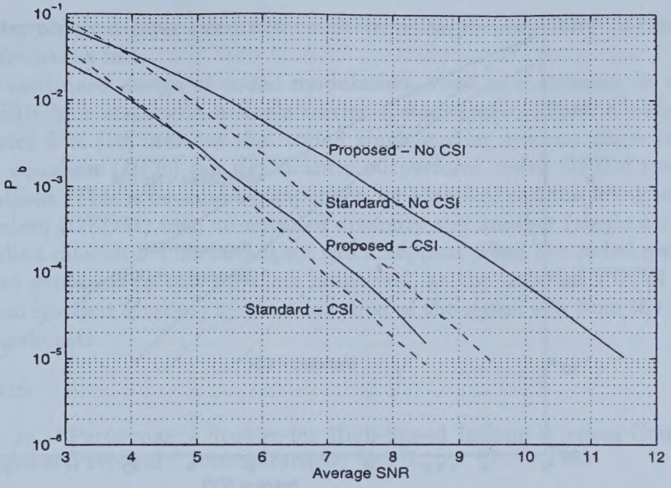


Figure 7: Bit error probability versus *average* signal-to-noise ratio per information bit for an OFDM system with a standard convolutional code ("Standard") and the proposed system, each with and without subchannel fading information ("CSI") at the receiver, when independent fading between the subcarriers carrying the dimensions of a signal drawn from S is assumed.

requires 2.8 to 3.6 dB more *average* signal-to-noise ratio to achieve $P_b = 10^{-4}$ than the standard scheme with CSI; thus, a gain of approximately 5.4 to 6.2 dB in required peak signal-to-noise ratio is still observed.

Before leaving this section, we comment on a couple of recent results not shown in Figures 7, 8, nor 9. First, if the operation of Figure 6 is removed (or, equivalently, let $\phi_j = 0, j = 0, \ldots, 3$) when CSI is employed at the receiver, only approximately 0.3 dB more average SNR is required to achieve $P_b = 10^{-5}$ over that required for the scheme of this section; furthermore, the removal of this operation results in a reduction in the number of extra complex multiplications per OFDM symbol required at the receiver from $4N$ to $2N$. Finally, further simulation results suggest that the optimal fixed phase modification, which maximizes the minimum product distance between signal vectors in S, can be decidely suboptimal (at least 0.5 dB at $P_b = 10^{-5}$) for the case when CSI is not available.

5 Conclusions

In this paper, the problem of designing coded modulation for peak-power constrained OFDM systems has been considered. By relaxing the restriction of subcarrier signal sets to traditional single-carrier signal sets, we are able to increase the diversity of the coded modulation scheme while controlling the envelope of the OFDM signal. Methods were developed for the design of coded modulation schemes, and these methods were employed to derive a simple coded modulation scheme which substantially outperforms standard coded BPSK schemes of approximately identical complexity in terms of required peak signal-to-noise ratio. The results indicate that compensation of the PMEPR problem

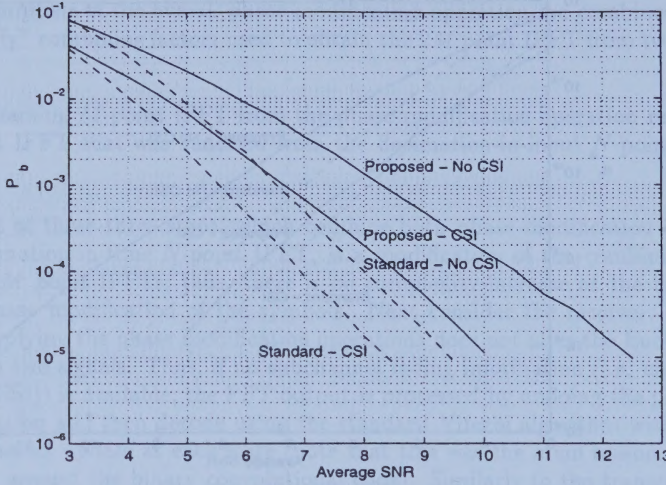


Figure 8: Bit error probability versus *average* signal-to-noise ratio per information bit for an OFDM system with a standard convolutional code ("Standard") and the proposed system, each with and without subchannel fading information ("CSI") at the receiver. The 6-tap GSM model for an urban area, $N = 16$, and a subchannel bandwidth of 100 KHz are employed.

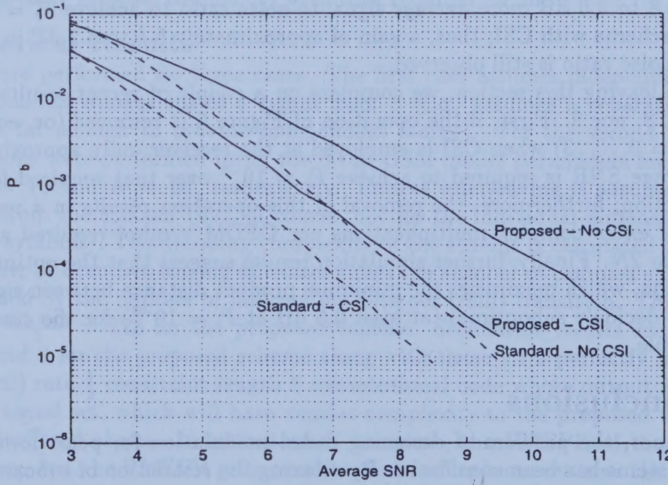


Figure 9: Bit error probability versus *average* signal-to-noise ratio per information bit for an OFDM system with a standard convolutional code ("Standard") and the proposed system, each with and without subchannel fading information ("CSI") at the receiver. The 6-tap GSM model for an urban area, $N = 64$, and a subchannel bandwidth of 100 KHz are employed.

of OFDM in the physical layer need not result in a large complexity increase nor a substantial performance loss.

Besides the continued design of coded modulation, work is continuing on the comparison of the FDCM system with a single-carrier system employing an equalizer. The results of Chapter 6 of [10] indicate that coded single-carrier systems employing a decision feedback equalizer (DFE) can significantly outperform coded OFDM systems for wireless applications. This is because a comparison for identical codes is being performed between one system (OFDM) with no uncoded diversity and another (single carrier with DFE) with implicit diversity. This advantage will disappear when the coded modulation schemes that are proposed in this work are employed, as the uncoded FDCM symbols are able to obtain implicit diversity by the relaxation of the signal sets from the standard single-carrier signal sets.

References

[1] L. Cimini, Jr., "Performance Studies for High-Speed Indoor Wireless Communications," *Wireless Personal Communications,* Vol. 2: pp. 67-85, 1995.

[2] T. Wilkinson and A. Jones, "Minimisation of the Peak to Mean Envelope Power Ratio of Multicarrier Transmission Schemes by Block Coding," *Proceedings of the 45^{th} Vehicular Technology Conference,* pp. 825-829, 1995.

[3] A. Jones and T. Wilkinson, "Combined Coding for Error Control and Increased Robustness to System Nonlinearities in OFDM," *Proceedings of the 46^{th} Vehicular Technology Conference,* pp. 904-908, 1996.

[4] R. van Nee, "OFDM Codes for Peak-to-Average Power Reduction and Error Correction," *Proceedings of the 1996 Global Communications Conference,* pp. 740-744, 1996.

[5] J. Davis and J. Jedwab, "Peak-to-Mean Power Control and Error Correction for OFDM Transmission using Golay Sequences and Reed-Muller Codes," *IEEE Electronics Letters,* Vol. 33, pp. 267-268, February 1997.

[6] S. Müller and J. Huber, "A Comparison of Peak Power Reduction Schemes for OFDM," *Proceedings of the 1997 Global Communications Conference,* pp. 1-5, 1997.

[7] K. Boulle and J. C. Belfiore, "Modulation Schemes Designed for the Rayleigh Channel," *Proceedings of the Conference on Information Sciences and Systems,* pp. 288-293, 1992.

[8] J. Boutros and E. Viterbo, "Signal Space Diversity: A Power- and Bandwidth-Efficient Diversity Technique for the Rayleigh Fading Channel," *IEEE Transactions on Information Theory,* pp. 1453-1467, July 1998.

[9] A. Oppenheim and R. Schafer, *Digital Signal Processing,* Englewood Cliffs: Prentice-Hall, 1975.

[10] S. Wilson, "Digital Audio Broadcasting in a Fading and Dispersive Channel," Ph.D. Dissertation, Stanford University, August 1994.

Profile Optimal 8-QAM and 32-QAM Constellations*

Xueting Liu and Richard D. Wesel
Electrical Engineering Department
University of California, Los Angeles
Los Angeles, CA 90095-1594
xueting@ee.ucla.edu and wesel@ee.ucla.edu

Abstract

For low-rate trellis codes that do not utilize uncoded bits, a new approach to constellation labeling which maximizes the constellation edge profile can provide improved metrics. We investigate the edge profile optimality for 8-QAM and 32-QAM constellations. The labeling structures that achieve the optimal edge profiles for 8-QAM and 32-QAM are presented. Explicit proofs of the optimal edge profiles for each case are demonstrated. Also several other quasi-Gray labelings for 32-QAM that are not profile optimal are given.

1 Introduction

For low-rate trellis codes that do not utilize uncoded bits, the standard techniques of set-partitioning and Gray labeling are not well motivated. A new approach to constellation labeling maximizes the constellations edge profile [1], [2]. This approach produces better trellis codes (in the sense of the broadcast code design metrics) than the standard techniques. A low-rate 32-QAM trellis code design example in [2] shows how the new approach can yield improved metrics at low rates.

This paper investigates 8-QAM and 32-QAM labeling strategies, identifying profile optimal labelings for each and proving their optimality. We also classify all quasi-Gray labelings for 32-QAM into profile optimal and non-profile optimal labelings.

Section 2 reviews the definition of the edge profile and some basic labeling lemmas [1] [2]. Section 3 presents the optimal edge profile 8-QAM constellation. Section 4 gives the structure and proof of optimal 32-QAM constellation. Section 5 provides quasi-Gray labelings for 32-QAM that are not profile superior. Conclusion is given in section 6, followed by an appendix containing proofs of several lemmas used in the paper.

2 Definition of the edge profile and basic labeling lemmas

The definition of the edge profile and three lemmas about labeled constellations appear in [2], [1]. Consider the constellation to be a fully connected graph with an **edge** between

---

*This research is supported with a grant from Pacific Bell in conjunction with UC Micro grant 98-169

each pair of constellation points. The binary symbol error that corresponds to an edge is its **edge label**. For a constellation with 2^n points, each binary symbol error labels 2^{n-1} edges. The edges with a given binary symbol error may not all have the same Euclidean distance. We are interested in the smallest distance associated with each edge label. We list definitions and basic lemmas below on which our following discussion are based.

Definition 1 *The edge profile of a labeled constellation is the monotonically increasing list of minimum squared distances, one for each nonzero edge label [2].*

Definition 2 *C_1 is profile superior to C_2 if the edge profile for C_1 is element by element greater than or equal to the edge profile for C_2, with at least one strict inequality [2].*

Lemma 1 *All the edges emanating from a given point must have different labels [1].*

Lemma 2 *Two different paths emanating from the same initial constellation point end at the same final constellation point and form a cycle iff they have the same edge label sum , i.e., the two different paths emanating from the same initial constellation point can't have the same edge label sum iff they form a new path instead of a cycle [1], [3].*

Lemma 3 *If all points in a constellation are connected by paths consisting only of edges with the minimum edge length, then the set of edge labels of the minimum length edges must contain a basis for the set of all edge labels [1].*

Another labeling lemma can be directly derived from Lemma 1 and Lemma 3.

Lemma 4 *If a basis has n distinct minimum-distance edge labels, then these basis vectors can label at most 2^n different constellation points.*

3 Optimal edge profile of 8-QAM

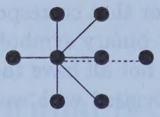
For the 8-QAM and 32-QAM constellations, since Gray codings are impossible [1], there are at least four distinct minimum-distance edge labels in 8-QAM and at least six distinct minimum-distance edge labels in 32-QAM. We refer to 8-QAM or 32-QAM constellation labelings with exactly four or six distinct minimum-distance edge labels as quasi-Gray coded. For convenience, we denote the minimum-distance as one unit-distance in the rest of the paper.

Figure 1(a) shows how to construct an edge profile upper bound for 8-QAM. By starting from a central constellation point and drawing an edge to every other constellation point, an upper bound can be formed by these edges since all these edges are distinct and the edge labels appear exactly once among them. Figure 1(b) gives the obtained upper bound for 8-QAM.

The labeling structures that achieve this bound are given in Figure 2 where $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{e}_z$ are linearly independent three-bit labels and $\mathbf{e}_4$ must be either $\mathbf{e}_4 = \mathbf{e}_x \oplus \mathbf{e}_y$ (possible only in (a)) or $\mathbf{e}_4 = \mathbf{e}_x \oplus \mathbf{e}_y \oplus \mathbf{e}_z$ (possible in (a), (b) and (c)).

4 Structure and proof of optimal 32-QAM

For the 32-QAM constellation, an upper bound and the best known achievable profile are given in Table 1 [2]. This upper bound is obtained by starting with the six distinct minimum distance edge labels and then using the same technique as 8-QAM.



(a)

E_i	1	2	3	4	5	6	7
$d^2_{\text{MIN}}(E_i)$	1	1	1	1	2	2	4

(b)

Figure 1: (a)Edges for profile bound in 8-QAM. (b)Edge profile bound for 8-QAM.

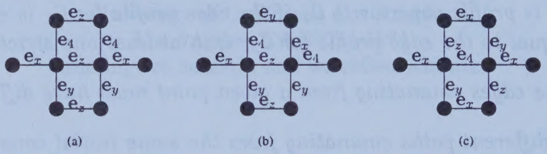


Figure 2: The optimal 8-QAM constellation labelings.

In this paper, the tightness of this upper bound is investigated and a tighter bound which is the same as the best known edge profile is given.

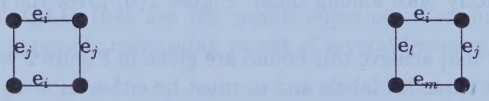
We begin with a lemma that partially characterizes the labeling structure imposed by six distinct minimum-distance edge labels.

Lemma 5 *In any valid quasi-Gray coded 32-QAM constellation labeling, one distinct minimum-distance edge label is equal to the XOR summation of three other distinct minimum-distance edge labels.*

Proof:
In any valid quasi-Gray coded 32-QAM constellation labeling, there are six distinct minimum-distance edge labels. Call these labels $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$, $\mathbf{e}_4$, $\mathbf{e}_5$ and $\mathbf{e}_6$ where $\mathbf{e}_1$ through $\mathbf{e}_5$ are the basis vectors. Then $\mathbf{e}_6$ is some linear combination of these basis vectors,

$$e_6 = k_1e_1 \oplus k_2e_2 \oplus k_3e_3 \oplus k_4e_4 \oplus k_5e_5\,, \quad \{k_i\} \in \{0,1\}. \tag{1}$$

According to labeling lemmas 1 and 2, the valid edge labels for a unit-distance square are shown in Figure 3 where $\{i,j,l,m\} \in \{1,2,3,4,5,6\}$. We will show that it is impossible to label a 32-QAM constellation only as in Figure 3(a). In other words, the structure shown in Figure 3(b) must appear in any valid 32-QAM constellation labeling.



(a) $\mathbf{e}_i$ and $\mathbf{e}_j$ are distinct.

(b) One is $\mathbf{e}_6$.

Figure 3: Valid edge labels in quasi-Gray coded 32-QAM.

Bound	1	1	1	1	1	1	2	2	4	4	4	4	5	5	5	5	5	5	5	5	8	8	8	9	9	10	10	10	10	13	13
Best known	1	1	1	1	1	1	2	2	2	2	2	2	2	2	4	4	5	5	5	5	5	5	5	5	8	8	8	10	10	13	13

Table 1: An upper bound for 32-QAM edge profiles and the best known achievable profile.

For a 32-QAM constellation, each row or column with six constellation points is labeled with at least three distinct unit-distance edge labels by Lemma 4. Now assuming we label 32-QAM only as in Figure 3(a) (denoted as Assumption 1), then all the horizontal edges in any column share the same label. Similarly, all the vertical edges in any row share the same label. Under Assumption 1, there are many types of labelings. In any such labeling, assume a minimum distance edge label to be a linear combination of any other five minimum distance edge labels. For any linear combination, a path (not a cycle) consisting of the edge label and its linear combination of edge labels can be found, which means this edge label is not equal to the corresponding linear combination by Lemma 2. This contradicts the fact that one of the six minimum-distance edge labels must be some linear combination of the basis vectors. Thus all such types of labeling structures under Assumption 1 are invalid, and therefore Figure 3 (b) must appear in any valid quasi-Gray coded 32-QAM labeling, i.e.,

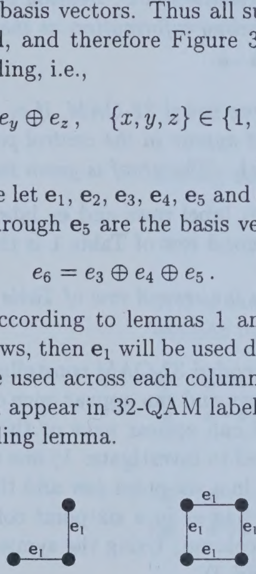
$$e_6 = e_x \oplus e_y \oplus e_z\,, \quad \{x, y, z\} \in \{1, 2, 3, 4, 5\}. \tag{2}$$

□

Without loss of generality, we let $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$, $\mathbf{e}_4$, $\mathbf{e}_5$ and $\mathbf{e}_6$ be the six distinct minimum-distance edge labels where $\mathbf{e}_1$ through $\mathbf{e}_5$ are the basis vectors and

$$e_6 = e_3 \oplus e_4 \oplus e_5\,. \tag{3}$$

Under these assumptions, according to lemmas 1 and 2, $\mathbf{e}_1$ ($\mathbf{e}_2$) can only label as Figure 3(a). If let one $\mathbf{e}_1$ label rows, then $\mathbf{e}_1$ will be used down each row. If also use one $\mathbf{e}_1$ to label columns, then $\mathbf{e}_1$ will be used across each column. Thus invalid subconstellation labelings as Figure 4 would then appear in 32-QAM labeling, in contradiction to Lemma 1. Thus we obtain another labeling lemma.



(a) (b)

Figure 4: Invalid subconstellation labeling if $\mathbf{e}_1$ labels both rows and columns.

Lemma 6 *In any valid quasi-Gray coded 32-QAM,* $\mathbf{e}_1$ ($\mathbf{e}_2$) *can't be used to label both rows and columns.*

From Lemma 5, we can derive additional lemmas that characterize the structure of edge profiles of quasi-Gray coded 32-QAM constellation labelings.

Lemma 7 *In any valid quasi-Gray coded 32-QAM constellation labeling, the structure of the edge profile must conform to Table 2 , assuming the minimum-distance is one unit-distance (The proof is given in the Appendix).*

The number of minimum-distance basis vectors in edge labels	Possible d^2_{MIN}	The number of edge labels
1	1	$\binom{5}{1} = 5$
2, 4	2, 4, 8, 10	$\binom{5}{2} + \binom{5}{4} = 15$
3, 5	1*, 5, 9, 13	$\binom{5}{3} + \binom{5}{5} = 11$

* 1 only possible here for the edge label $\mathbf{e}_6 = \mathbf{e}_3 \oplus \mathbf{e}_4 \oplus \mathbf{e}_5$

Table 2: The edge profile structure of a valid quasi-Gray coded 32-QAM constellation labeling.

Lemma 8 *In any valid quasi-Gray coded 32-QAM,* $\mathbf{e}_1$ *and* $\mathbf{e}_2$ *can't both be used to label columns or both be used to label rows. (The proof is given in the Appendix).*

Lemma 9 *When two* $\mathbf{e}_1$ *or two* $\mathbf{e}_2$ *appear in a horizontal or vertical path, an equivalent collapsed version of the subconstellation (i.e., treating the two points connected by* $\mathbf{e}_1$ *or* $\mathbf{e}_2$ *as one point) contains all labeling information as the original subconstellation. (The proof is given in the Appendix).*

Lemma 10 *In a valid quasi-Gray coded 32-QAM, if* $\mathbf{e}_1$ *(*$\mathbf{e}_2$*) appears once in a six-point row (column), then* $\mathbf{e}_1$ *(*$\mathbf{e}_2$*) must appear in the central position, and* $\mathbf{e}_1$ *(*$\mathbf{e}_2$*) can't appear twice in a six-point row (column). (The proof is given in the Appendix).*

Without loss of generality, let $\mathbf{e}_1$ label rows and $\mathbf{e}_2$ label columns. We now prove that the edge length profile in the second row of Table 1 is the optimal edge profile.

Theorem 1 *The edge profile in the second row of Table 1 is the optimal edge profile for quasi-Gray coded 32-QAM.*

Proof: In any valid quasi-Gray coded 32-QAM constellation, according to Lemmas 1, 2, 6, 8 and 10, $\mathbf{e}_1$ is used to label rows and can appear once or three times in a six-point row; $\mathbf{e}_2$ is used to label columns and can appear once or three times in a six-point column. Thus there are four cases we need to investigate: 1) one $\mathbf{e}_1$ in a six-point row and one $\mathbf{e}_2$ in a six-point column, 2) one $\mathbf{e}_1$ in a six-point row and three $\mathbf{e}_2$ in a six-point column, 3) three $\mathbf{e}_1$ in a six-point row and one $\mathbf{e}_2$ in a six-point column, 4) three $\mathbf{e}_1$ in a six-point row and three $\mathbf{e}_2$ in a six-point column. Using the symmetry of $\mathbf{e}_1$ and $\mathbf{e}_2$, we only need to consider the cases of 1), 2) and 4).

Case 1: $\mathbf{e}_1$ appears once in a six-point row and $\mathbf{e}_2$ appears once in a six-point column.

The structure of 32-QAM for Case 1 is shown in Figure 5(a). First consider the labelings of the subconstellation circled by dotted lines in Figure 5(a) where $\mathbf{e}_x \neq \mathbf{e}_y$, $\mathbf{e}_y \neq \mathbf{e}_j$, $\mathbf{e}_j \neq \mathbf{e}_k$ and $\{x, y, j, k\} \in \{3, 4, 5, 6\}$ by Lemma 1. According to Lemma 4, at least three distinct basis vectors must be from $\mathbf{e}_x$, $\mathbf{e}_y$, $\mathbf{e}_j$ and $\mathbf{e}_k$. By Equation 3, $\mathbf{e}_x$, $\mathbf{e}_y$, $\mathbf{e}_j$ and $\mathbf{e}_k$ can't be four distinct basis vectors and among $\mathbf{e}_x$, $\mathbf{e}_y$, $\mathbf{e}_j$ and $\mathbf{e}_k$, one of them must be equal to another one. By following the labeling lemmas, the three possible types of labeling for the subconstellation are shown in Figure 5(b)-(d).

Using the symmetry of $\mathbf{e}_1$ and $\mathbf{e}_2$, figures 5(c) and (d) are isomorphic. Thus we only need to consider the case of figures 5(b) and (c).

Figure 6 shows the two possible labeling structures for Case 1, where $\mathbf{e}_j$, $\mathbf{e}_k$, $\mathbf{e}_y$ and $\mathbf{e}_m$ are distinct, {j, k, y, m} ∈ {3, 4, 5, 6} and $\mathbf{e}_m = \mathbf{e}_j \oplus \mathbf{e}_k \oplus \mathbf{e}_y$. From figures 6(a) and (b), we can obtain Table 3 which represents the upper bound of edge profile of Case 1.

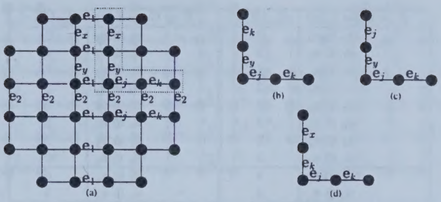
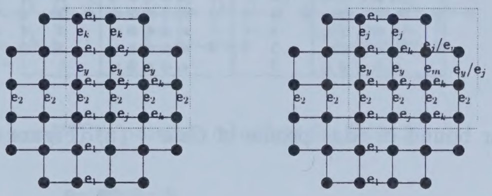


Figure 5: (a) 32QAM labeling structure of Case 1. (b), (c), (d) Subconstellation labelings of Case 1.



(a) (b)

Figure 6: Two possible kinds of labeling structures for Case 1.

Case 2: $\mathbf{e}_1$ appears once in a six-point row and $\mathbf{e}_2$ appears three times in a six-point column.

For this case, the labeling structure of 32-QAM is shown in Figure 7 where $\mathbf{e}_x \neq \mathbf{e}_y \neq \mathbf{e}_j$. First consider the labelings of the subconstellation inside the dotted rectangular box.

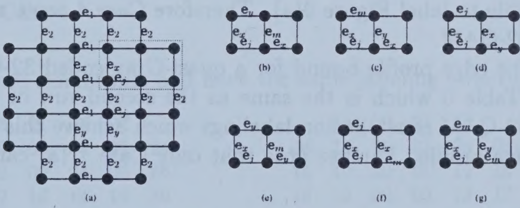


Figure 7: (a) 32QAM labeling structure of Case 2. (b)-(g) six possible labelings for the subconstellation of Case 2.

Six possible labelings for this subconstellation are shown in figures 7(b)-(g) where $\mathbf{e}_j$, $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{e}_m$ are distinct, {j, x, y, m} ∈ {3, 4, 5, 6} and $\mathbf{e}_m = \mathbf{e}_j \oplus \mathbf{e}_x \oplus \mathbf{e}_y$. Correspondingly there are six possible constellation labelings which are shown in figures 8(a)-(f). Using the labeling lemmas, from figures 8(a)-(f), we obtain Table 4, which represents the upper bound of the constellation labelings respectively where figures 8(a)(c) are isomorphic and so are figures 8(b)(d).

Using Lemma 7, Table 3 and Table 4, the upper bound of the edge profiles of Case 1(a)(b)and Case 2(a)-(f) can be obtained. We put the upper bound of Case 1 and Case 2 in Table 5 where all these upper bounds of edge profiles are dominated by Case 1(a).

E *in Case* 1 (a)	$d^2_{\text{MIN}}(E) \leq$	E *in Case* 1 (b)	$d^2_{\text{MIN}}(E) \leq$
$\mathbf{e}_1 \oplus \mathbf{e}_2$	2	$\mathbf{e}_1 \oplus \mathbf{e}_2$	2
$\mathbf{e}_1 \oplus \mathbf{e}_k$	2	$\mathbf{e}_1 \oplus \mathbf{e}_j$	2
$\mathbf{e}_1 \oplus \mathbf{e}_y$	2	$\mathbf{e}_1 \oplus \mathbf{e}_y$	2
$\mathbf{e}_2 \oplus \mathbf{e}_j$	2	$\mathbf{e}_2 \oplus \mathbf{e}_j$	2
$\mathbf{e}_2 \oplus \mathbf{e}_k$	2	$\mathbf{e}_2 \oplus \mathbf{e}_k$	2
$\mathbf{e}_k \oplus \mathbf{e}_j$	2	$\mathbf{e}_k \oplus \mathbf{e}_j$	2
$\mathbf{e}_k \oplus \mathbf{e}_y$	2	$\mathbf{e}_k \oplus \mathbf{e}_y$	2
$\mathbf{e}_y \oplus \mathbf{e}_j$	2	$\mathbf{e}_y \oplus \mathbf{e}_j$	2
$\mathbf{e}_1 \oplus \mathbf{e}_j$	4	$\mathbf{e}_1 \oplus \mathbf{e}_k$	4
$\mathbf{e}_2 \oplus \mathbf{e}_y$	4	$\mathbf{e}_2 \oplus \mathbf{e}_y$	4
$\mathbf{e}_y \oplus \mathbf{e}_2 \oplus \mathbf{e}_1 \oplus \mathbf{e}_j$	8	$\mathbf{e}_2 \oplus \mathbf{e}_m$	4
$\mathbf{e}_1 \oplus \mathbf{e}_k \oplus \mathbf{e}_y \oplus \mathbf{e}_j$	8	$\mathbf{e}_y \oplus \mathbf{e}_2 \oplus \mathbf{e}_1 \oplus \mathbf{e}_j$	8
$\mathbf{e}_2 \oplus \mathbf{e}_k \oplus \mathbf{e}_y \oplus \mathbf{e}_j$	8	$\mathbf{e}_2 \oplus \mathbf{e}_1 \oplus \mathbf{e}_y \oplus \mathbf{e}_k$	8
		$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_y \oplus \mathbf{e}_k$	8
$\mathbf{e}_1 \oplus \mathbf{e}_y \oplus \mathbf{e}_k$	5	$\mathbf{e}_1 \oplus \mathbf{e}_y \oplus \mathbf{e}_k$	5
$\mathbf{e}_1 \oplus \mathbf{e}_y \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_y \oplus \mathbf{e}_j$	5
$\mathbf{e}_1 \oplus \mathbf{e}_j \oplus \mathbf{e}_k$	5	$\mathbf{e}_1 \oplus \mathbf{e}_j \oplus \mathbf{e}_k$	5
$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_k$	5	$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_k$	5
$\mathbf{e}_2 \oplus \mathbf{e}_y \oplus \mathbf{e}_k$	5	$\mathbf{e}_2 \oplus \mathbf{e}_y \oplus \mathbf{e}_k$	5
$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_y$	5	$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_y$	5
$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	5
$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_j$	5

Table 3: The upper bound of edge profile of Case 1 (a) (Figure 6(a)) and Case 1 (b) (Figure 6(b)).

All these upper bounds of edge profiles are achievable. The labelings which achieve these upper bounds will be discussed in the next section.

Case 4: $\mathbf{e}_1$ appears three times in a six-point row and $\mathbf{e}_2$ appears three times in a six-point column.

Figure 9 shows the 32-QAM labeling structure for Case 4. We apply Lemma 9 to get the equivalent subconstellation. All other blank edges must be labeled by using $\mathbf{e}_3$ through $\mathbf{e}_6$.

In Figure 9(b), we need to label 9 different points by using $\mathbf{e}_3$ through $\mathbf{e}_6$, which have only three basis vectors. By Lemma 4, it is impossible to label Figure 9(b). Correspondingly it is not possible to label Figure 9(a). Therefore Case 4 can't appear in any valid quasi-Gray coded 32-QAM.

In conclusion, the edge profile bound for a quasi-Gray coded 32-QAM constellation is the first row of Table 5 which is the same as the second row of Table 1. All valid quasi-Gray coded 32-QAM constellation labelings which achieve this edge profile bound have an optimal edge profile. Further note that only Case 1 (a) can achieve this edge profile bound. □

5 Six distinct classes of Quasi-Gray labeled 32-QAM

The labeling [2] which achieves the optimal edge profile is shown below in octal (on the left-hand-side). All other labelings which achieve the optimal edge profile are isomorphic and have the same labeling structure as Case 1(a) (illustrated in Figure 6(a)). Several other quasi-Gray labelings that are not profile optimal exist. Corresponding to the labeling structure of Case 1(b) (illustrated in Figure 6 (b)), one labeling which achieves this profile bound is shown below in octal (on the right hand side). All other labelings

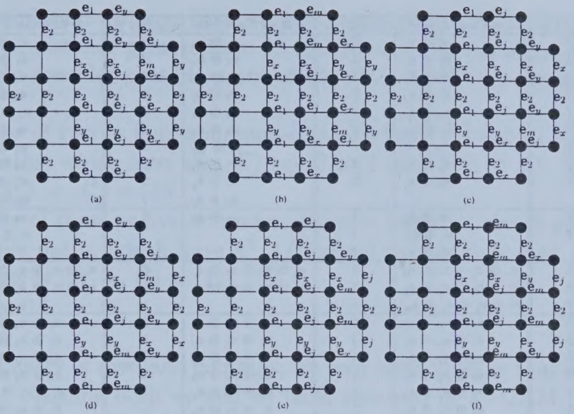
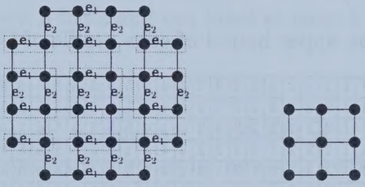


Figure 8: Six possible labelings for Case 2.



(a) (b)

Figure 9: (a) The labeling structure for Case 4. (b) The equivalent subconstellation labeling structure for Case 4.

achieving this profile are isomorphic and have the same labeling structure as Case 1(b).

	23	21	25	27	
20	03	01	05	07	24
22	02	00	04	06	26
32	12	10	14	16	36
30	13	11	15	17	34
	33	31	35	37	

	34	30	31	35	
04	24	20	21	25	05
14	10	00	01	11	15
16	12	02	03	13	17
06	26	22	23	27	07
	36	32	33	37	

Corresponding to Case 2(a)(c), Case 2(b)(d), Case 2(e) and Case 2(f), the labelings which achieve their respective upper bounds of edge profiles are given below from left hand side to right hand side. Furthermore they have isomorphic labelings and share the same labeling structures given in our proof.

	16	06	07	17	
34	14	04	05	15	35
24	20	00	01	21	25
26	22	02	03	23	27
36	32	12	13	33	37
	30	10	11	31	

	32	06	07	33	
34	30	04	05	31	35
24	20	00	01	21	25
26	22	02	03	23	27
36	16	12	13	17	37
	14	10	11	15	

	26	06	07	27	
34	24	04	05	25	35
14	20	00	01	21	15
16	22	02	03	23	17
36	32	12	13	33	37
	30	10	11	31	

	26	12	13	27	
34	24	10	11	25	35
30	04	00	01	05	31
32	06	02	03	07	33
36	16	22	23	17	37
	14	20	21	15	

E in Case 2 (a)(c)	$d^2_{\text{MIN}}(E) \le$	*E in Case 2 (b)(d)*	$d^2_{\text{MIN}}(E) \le$	*E in Case 2 (e)*	$d^2_{\text{MIN}}(E) \le$	*E in Case 2 (f)*	$d^2_{\text{MIN}}(E) \le$
$\mathbf{e}_1 \oplus \mathbf{e}_2$	2	$\mathbf{e}_1 \oplus \mathbf{e}_2$	2	$\mathbf{e}_1 \oplus \mathbf{e}_2$	2	$\mathbf{e}_1 \oplus \mathbf{e}_2$	2
$\mathbf{e}_1 \oplus \mathbf{e}_x$	2	$\mathbf{e}_1 \oplus \mathbf{e}_x$	2	$\mathbf{e}_1 \oplus \mathbf{e}_x$	2	$\mathbf{e}_1 \oplus \mathbf{e}_x$	2
$\mathbf{e}_1 \oplus \mathbf{e}_y$	2	$\mathbf{e}_1 \oplus \mathbf{e}_y$	2	$\mathbf{e}_1 \oplus \mathbf{e}_y$	2	$\mathbf{e}_1 \oplus \mathbf{e}_y$	2
$\mathbf{e}_2 \oplus \mathbf{e}_j$	2	$\mathbf{e}_2 \oplus \mathbf{e}_j$	2	$\mathbf{e}_2 \oplus \mathbf{e}_j$	2	$\mathbf{e}_2 \oplus \mathbf{e}_j$	2
$\mathbf{e}_2 \oplus \mathbf{e}_x$	2	$\mathbf{e}_2 \oplus \mathbf{e}_x$	2	$\mathbf{e}_2 \oplus \mathbf{e}_x$	2	$\mathbf{e}_2 \oplus \mathbf{e}_m$	2
$\mathbf{e}_2 \oplus \mathbf{e}_y$	2	$\mathbf{e}_2 \oplus \mathbf{e}_m$	2	$\mathbf{e}_2 \oplus \mathbf{e}_y$	2	$\mathbf{e}_2 \oplus \mathbf{e}_x$	2
$\mathbf{e}_y \oplus \mathbf{e}_j$	2	$\mathbf{e}_y \oplus \mathbf{e}_j$	2	$\mathbf{e}_2 \oplus \mathbf{e}_m$	2	$\mathbf{e}_2 \oplus \mathbf{e}_y$	2
$\mathbf{e}_x \oplus \mathbf{e}_y$	2	$\mathbf{e}_x \oplus \mathbf{e}_y$	2	$\mathbf{e}_y \oplus \mathbf{e}_j$	2	$\mathbf{e}_y \oplus \mathbf{e}_j$	2
$\mathbf{e}_x \oplus \mathbf{e}_j$	2	$\mathbf{e}_x \oplus \mathbf{e}_j$	2	$\mathbf{e}_x \oplus \mathbf{e}_y$	2	$\mathbf{e}_x \oplus \mathbf{e}_j$	2
				$\mathbf{e}_x \oplus \mathbf{e}_j$	2	$\mathbf{e}_x \oplus \mathbf{e}_y$	2
$\mathbf{e}_1 \oplus \mathbf{e}_j$	4	$\mathbf{e}_1 \oplus \mathbf{e}_j$	4	$\mathbf{e}_1 \oplus \mathbf{e}_j$	4	$\mathbf{e}_1 \oplus \mathbf{e}_j$	4
$\mathbf{e}_2 \oplus \mathbf{e}_m$	4	$\mathbf{e}_1 \oplus \mathbf{e}_m$	4			$\mathbf{e}_1 \oplus \mathbf{e}_m$	4
		$\mathbf{e}_2 \oplus \mathbf{e}_y$	4				
$\mathbf{e}_x \oplus \mathbf{e}_2 \oplus \mathbf{e}_1 \oplus \mathbf{e}_j$	8	$\mathbf{e}_x \oplus \mathbf{e}_2 \oplus \mathbf{e}_1 \oplus \mathbf{e}_j$	8	$\mathbf{e}_x \oplus \mathbf{e}_2 \oplus \mathbf{e}_1 \oplus \mathbf{e}_j$	8	$\mathbf{e}_x \oplus \mathbf{e}_2 \oplus \mathbf{e}_1 \oplus \mathbf{e}_j$	8
$\mathbf{e}_j \oplus \mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	8	$\mathbf{e}_j \oplus \mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	8	$\mathbf{e}_j \oplus \mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	8	$\mathbf{e}_j \oplus \mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	8
$\mathbf{e}_y \oplus \mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_x$	8	$\mathbf{e}_1 \oplus \mathbf{e}_j \oplus \mathbf{e}_2 \oplus \mathbf{e}_m$	8				
$\mathbf{e}_1 \oplus \mathbf{e}_x \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_x \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_x \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_x \oplus \mathbf{e}_j$	5
$\mathbf{e}_1 \oplus \mathbf{e}_y \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_y \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_y \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_y \oplus \mathbf{e}_j$	5
$\mathbf{e}_1 \oplus \mathbf{e}_x \oplus \mathbf{e}_y$	5	$\mathbf{e}_1 \oplus \mathbf{e}_x \oplus \mathbf{e}_y$	5	$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_y$	5	$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_y$	5
$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_y$	5	$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_y$	5	$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_x$	5	$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_x$	5
$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_x$	5	$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_x$	5	$\mathbf{e}_2 \oplus \mathbf{e}_x \oplus \mathbf{e}_y$	5	$\mathbf{e}_2 \oplus \mathbf{e}_j \oplus \mathbf{e}_m$	5
$\mathbf{e}_2 \oplus \mathbf{e}_x \oplus \mathbf{e}_y$	5	$\mathbf{e}_2 \oplus \mathbf{e}_x \oplus \mathbf{e}_y$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	5
$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_y$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_x$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_x$	5
$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_x$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_x$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_j$	5
$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_j$	5	$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_j$	5			$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_m$	5
		$\mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_m$	5				
				$\mathbf{e}_1 \oplus \mathbf{e}_j \oplus \mathbf{e}_m$	9	$\mathbf{e}_1 \oplus \mathbf{e}_j \oplus \mathbf{e}_m$	9

Table 4: The upper bound of edge profile of Case 2.

Case 1 (a)	1	1	1	1	1	1	2	2	2	2	2	2	2	2	4	4	5	5	5	5	5	5	5	5	8	8	8	10	10	13	13
Case 1 (b)	1	1	1	1	1	1	2	2	2	2	2	2	2	2	4	4	4	5	5	5	5	5	5	5	5	8	8	8	10	13	13
Case 2 (a), (c)	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	4	4	5	5	5	5	5	5	5	5	5	8	8	8	10	13
Case 2 (b), (d)	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	4	4	4	5	5	5	5	5	5	5	5	5	5	8	8	8
Case 2 (e)	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	4	5	5	5	5	5	5	5	5	8	8	9	10	10	13
Case 2 (f)	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	4	4	5	5	5	5	5	5	5	5	5	8	8	9	10

Table 5: The upper bounds for 32-QAM edge profiles of Case 1 and Case 2.

6 Conclusion

We have identified the optimal edge profile structures for both 8-QAM and Quasi-Gray coded 32-QAM constellations. Several other quasi-Gray labeling structures for 32-QAM that are not profile optimal also have been given. For low rate codes, a better edge profile provides larger Euclidean distance which is demonstrated in a code design example by wesel [2]. We also conjecture that even high rate trellis codes benifit from a superior edge profile when the number of states is large enough. But this is a moot point, since considering the level of complexity, it is more worthwhile to use turbo codes than trellis codes.

Appendix

Proof of Lemma 7

Proof:

By Equation 3, some edge labels having two minimum-distance edge labels can be treated as having four minimum-distance edge labels and vice versa. Similarly some edge labels having three minimum-distance edge labels can be treated as having five minimum-distance edge labels and vice versa. Thus we group the cases of having two or four minimum-distance basis vectors in edge labels together and group the cases of having three or five minimum-distance basis vectors in edge labels together. Since no edge label has the squared distance greater than 13 in 32-QAM edge profiles shown in the bound, we only need to consider possible paths of edge labels with the squared distances less or equal to 13.

For the case of two or four minimum-distance basis vectors in edge labels, there are six possible paths, which have $d^2_{\text{MIN}}(E)$ equal to 2 or 4 or 8 or 10. The number of such edge labels is $\binom{5}{2} + \binom{5}{4} = 15$.

For the case of three or five minimum-distance basis vectors in edge labels, there are also six possible paths, which have $d^2_{\text{MIN}}(E)$ equal to 1 or 5 or 9 or 13. The number of such edge labels is $\binom{5}{3} + \binom{5}{5} = 11$. □

Proof of Lemma 8

Proof:

Using the symmetry of rows and columns, we only need to show that $\mathbf{e}_1$ and $\mathbf{e}_2$ can't both label columns. Assume both $\mathbf{e}_1$ and $\mathbf{e}_2$ label columns in 32-QAM, then they can't be used to label rows according to Lemma 6. Since each column with six constellation points must be labeled with at least three distinct minimum-distance edge labels by Lemma 4, then there is at least one column that is not labeled with either $\mathbf{e}_1$ or $\mathbf{e}_2$. Therefore we have to label the subconstellations with 12 points or 10 points by using $\mathbf{e}_3$, $\mathbf{e}_4$, $\mathbf{e}_5$ and $\mathbf{e}_6$ (only three of them form a basis and can label at most 8 points), which is impossible. □

Proof of Lemma 9

Proof:

This lemma is true because $\mathbf{e}_1$, $\mathbf{e}_2$ can only be labeled as in Figure 3 (a). So by applying the collapse rule and Lemma 2, the labeling information would not get lost in the collapsed version. □

Proof of Lemma 10

Proof:

Using the symmetry of $\mathbf{e}_1$ and $\mathbf{e}_2$, rows and columns, we only need to consider the case of $\mathbf{e}_1$ appearing in rows.

If $\mathbf{e}_1$ labels only one edge label and does not appear in the central position, then we have to label subconstellations with 28 or 22 points by using $\mathbf{e}_2$, $\mathbf{e}_3$, $\mathbf{e}_4$, $\mathbf{e}_5$ and $\mathbf{e}_6$ (only four of them form a basis and can label at most 16 points), which is impossible. Thus $\mathbf{e}_1$ must appear in the central position if $\mathbf{e}_1$ appears only once.

If $\mathbf{e}_1$ appears twice in a six-point row, there are five possible labeling structures. Among these five possible labeling structures, we have to label either subconstellations with 18 or 24 points or collapsed subconstellations with 18 or 20 points all by using $\mathbf{e}_2$ through $\mathbf{e}_6$ (only four of them form a basis and can label at most 16 points), which is impossible. Therefore $\mathbf{e}_1$ can't appear twice in a six-point row in a valid quasi-Gray coded 32-QAM. □

References

[1] R. D. Wesel. *Trellis Code Design for Correlated Fading and Achievable Rates for Tomlinson Harashima Precoding.* PhD thesis, Stanford University, Aug. 1996.

[2] R. D. Wesel, C. Komninakis, and X. Liu. *Towards Optimality in Constellation Labeling.* In *Proc. of Globecom '97*, Phoenix, AZ, November 3-8 1997.

[3] F. Harary. *Graph Theory.* Addison Wesley Longman, Incorporated, 1994.

MULTICHANNEL RECEPTION OF DIGITAL SIGNALS OVER CORRELATED NAKAGAMI FADING CHANNELS

Mohamed-Slim Alouini

Dept. of Electrical & Computer Eng.
University of Minnesota
4-174 EE/CSci Building
200 Union Street S. E.
Minneapolis, MN 55455, USA.
E-mail : `<alouini@ece.umn.edu>`

Marvin K. Simon

Jet Propulsion Laboratory
California Institute of Technology
Mail Stop: 238-343
400 Oak Grove Drive
Pasadena, CA 91109-8099, USA.
E-mail: `<marvin.k.simon@jpl.nasa.gov>`

ABSTRACT

We obtain generic expressions for the average error rate of multichannel reception of digital signals over correlated slowly varying Nakagami-m fading channels. The results are applicable to coherent detection with maximal-ratio combining as well as to differentially coherent and noncoherent detection with post-detection equal-gain combining. Aside from extending previous analyses of diversity reception over correlated Nakagami-m fading channels to include the performance of M-ary modulations, these results provide equivalent forms for known expressions corresponding to the performance of binary modulations. Because of their simple forms, these results offer a useful analytical tool for the accurate performance evaluation of various systems of practical interest.

1. INTRODUCTION

In studying the performance of diversity systems, the usual assumption made is that the combined signals are independent of one another [1, 2]. As discussed in [3, 4, 5, 6, 7, 8] there are a number of real life scenarios in which this assumption is not valid because, for example, of insufficient antenna spacing in small-size mobile units equipped with space antenna diversity. As a result, the maximum theoretical diversity gain cannot be achieved and hence one must revamp their analysis to account for the effect of correlation between the combined signals. Along these lines, several correlation models have been proposed [3, 6, 9], and using these models several authors [3, 4, 5, 6, 7, 8] have analyzed special cases of the performance of various systems corresponding to specific detection, modulation, and diversity combining schemes. For instance, Pierce and Stein [3] considered the performance of binary coherent and noncoherent systems over correlated identically distributed Rayleigh fading channels. In particular they obtained the average bit error rate (BER) of coherent binary phase-shift-keying (BPSK) when used in conjunction with maximal-ratio combining (MRC) [10], [11, Section 5.5.3, p. 24], and of noncoherent frequency-shift-keying (BFSK) when used with postdetection equal-gain combining (EGC) [11, Section 5.5.6, p. 253], [12, Section 4.4, p. 298]. Miyagaki *al.* [4] analyzed the outage probability and the average symbol error rate (SER) performance of M-ary phase-shift-keying (M-PSK) for various dual-branch diversity receivers over correlated identically distributed Nakagami-m fading channels [9]. Al-Hussaini and Al-Bassiouni [5] obtained a closed form expression for the average BER of noncoherent BFSK with dual-branch MRC reception over correlated non identically distributed Nakagami-m fading channels. More recently Aalo [6] analyzed the outage probability and the average BER of various coherent, differentially coherent, and noncoherent binary modulations with multichannel MRC reception over identically distributed Nakagami-m fading channels with two correlation models, namely the constant (equal) correlation model and the exponential correlation model. Finally, Patenaude *et al.* [7, 8] extended the results of Al-Hussaini and Al-Bassiouni and of by providing closed form expressions for the average BER of orthogonal noncoherent BFSK with postdetection EGC reception over two correlated nonidentical and also D equicorrelated identically distributed Nakagami-m channels.

In this paper we obtain general results for the exact average BER or SER of M-ary coherent, differentially-coherent, and noncoherent modulations over equicorrelated, exponentially, and arbitrary correlated slowly varying Nakagami-m fading channels. Aside from allowing for many detection, modulation, and diversity combining cases not previously treated, these general results provide often simpler forms for the average BER or SER expressions corresponding to the special cases treated by the cited authors. In particular, each average BER or SER expression is obtained in the form of a single integral with finite limits and an integrand composed of elementary functions.

The remainder of this paper is organized as follows. In the next section, the various channel correlation models under consideration are described. Section 3 provides in a generic fashion the performance of coherent modulations with MRC reception. Section 4 extends these performance results to differentially coherent and noncoherent modulations with postdetection EGC reception. Due to space limitations, numerical examples for the error rates of particular modulations of interest over various correlated fading channels are omitted here but will be presented in an extended version of this paper [13] which in addition will cover some applications of the framework presented in this paper to the performance evaluation of hybrid diversity systems (such as space-multipath or frequency-multipath systems) over correlated fading channels.

2. CHANNEL CORRELATION MODELS

In this section we present the four channel correlation models (models A, B, C, and D) of interest in this paper. Since both of the diversity schemes (MRC and postdetection EGC) under consideration are known to have conditional error rates which are function of the sum of the combined individual signals' signal-to-noise ratio (SNR) per symbol, $\gamma_d = r_d^2 E_s/N_0$ $(d = 1, 2, \cdots, D)$, where d is the branch index, D is the number of combined branches, r_d is the dth branch fading amplitude, and E_s/N_0 is the symbol energy-to-Gaussian noise spectral density ratio, we present for every model the probability function density (PDF) for the combined SNR per symbol, $\gamma_t = \sum_{d=1}^{D} \gamma_d$. Furthermore since our generic approach to evaluate the average error rate performance relies on the knowledge of the moment generating function (MGF) of the combined SNR PDF we derive this quantity for the four channel models under consideration. In what follows, $p_i(\gamma_t)$ denotes the PDF of the combined SNR per symbol, with the index i identifying the model type and is hence equal to a, b, c, or d. In addition, the MGF of $p_i(\gamma_t)$ is denoted by $\mathcal{M}_i(s)$ and is defined by

$$\mathcal{M}_i(s) \triangleq E_{\gamma_t}\left[e^{s\gamma_t}\right] = \int_0^\infty p_i(\gamma_t)\, e^{s\gamma_t}\, d\gamma_t. \tag{1}$$

2.1. Model A: Two Correlated Branches with Nonidentical Fading

Model A was proposed by Nakagami [9, Section 6.4] and corresponds to the scenarios of dual diversity reception over correlated Nakagami-m channels which are not necessarily identically distributed.

2.1.1. PDF

In this case the PDF of the combined signal envelope, $p_a(r_t)$, is given by [9, Eq. (142), p.34]

$$p_a(r_t) = \frac{2r_t\sqrt{\pi}}{\Gamma(m)(\sigma_1\sigma_2(1-\rho))^m}\left(\frac{r_t^2}{2\beta}\right)^{m-\frac{1}{2}} I_{m-\frac{1}{2}}(\beta r_t^2)\, e^{-\alpha r_t^2};\quad r_t \geq 0, \tag{2}$$

where $I_\nu(\cdot)$ denotes the νth-order modified Bessel function [14, Section10.2, p. 443],

$$\rho = \frac{\operatorname{cov}(r_1^2, r_2^2)}{\sqrt{\operatorname{var}(r_1^2)\operatorname{var}(r_2^2)}},\quad 0 \leq \rho < 1. \tag{3}$$

is the envelope correlation coefficient between the two signals [1], and the parameters σ_d (d=1, 2), α, and β are defined as follows:

$$\sigma_d = \frac{\Omega_d}{m},\quad (d = 1, 2), \tag{4}$$

$$\alpha = \frac{(\sigma_1 + \sigma_2)}{2\sigma_1\sigma_2(1-\rho)}, \tag{5}$$

$$\beta^2 = \frac{(\sigma_1 - \sigma_2)^2 + 4\sigma_1\sigma_2\rho}{4\sigma_1^2\sigma_2^2(1-\rho)^2}. \tag{6}$$

By using a standard transformation of random variables, it can be shown that the PDF of the combined SNR per symbol, $p_a(\gamma_t)$, is given by

$$p_a(\gamma_t) = \frac{\sqrt{\pi}}{\Gamma(m)}\left[\frac{m^2}{\overline{\gamma}_1\overline{\gamma}_2(1-\rho)}\right]^m \left(\frac{\gamma_t}{2\beta'}\right)^{m-\frac{1}{2}} I_{m-\frac{1}{2}}(\beta'\gamma_t)\, e^{-\alpha'\gamma_t};\quad \gamma_t \geq 0, \tag{7}$$

where the parameters α' and β' are normalized version of the parameters α and β, and are given by

$$\alpha' \triangleq \frac{\alpha}{E_s/N_0} = \frac{m(\overline{\gamma}_1 + \overline{\gamma}_2)}{2\overline{\gamma}_1\overline{\gamma}_2(1-\rho)}, \tag{8}$$

$$\beta' \triangleq \frac{\beta}{E_s/N_0} = \frac{m\left((\overline{\gamma}_1 + \overline{\gamma}_2)^2 - 4\overline{\gamma}_1\overline{\gamma}_2(1-\rho)\right)^{1/2}}{2\overline{\gamma}_1\overline{\gamma}_2(1-\rho)}. \tag{9}$$

For the cases of identical channels ($\overline{\gamma}_1 = \overline{\gamma}_2 = \overline{\gamma}$), (7) reduces to

$$p_a(\gamma_t) = \frac{\pi}{\Gamma(m)\sqrt{1-\rho}}\left(\frac{m}{\overline{\gamma}}\right)^{m+\frac{1}{2}}\left(\frac{\gamma_t}{2\sqrt{\rho}}\right)^{m-\frac{1}{2}} \exp\left[-\frac{m\gamma_t}{(1-\rho)\overline{\gamma}}\right] I_{m-\frac{1}{2}}\left(\frac{m\sqrt{\rho}\gamma_t}{\overline{\gamma}(1-\rho)}\right);\quad \gamma_t \geq 0, \tag{10}$$

For the Rayleigh fading case ($m = 1$) using the identity [14, Eq. (10.2.13), p.443]

$$\sqrt{\frac{\pi}{2z}}\, I_{\frac{1}{2}}(z) = \frac{\sinh z}{z}, \tag{11}$$

---

[1] We use the envelope correlation coefficient throughout this paper as a measure of the degree of correlation between the fading signals since, as pointed it out in [3, 6], experimental data on the correlation between fading signals are typically given in terms of this figure because of its relative ease of measurement.

where sinh(·) denotes the sin hyperbolic function [14, Section 4.5, p.83], it can be shown that (7) reduces to

$$p_a(\gamma_t) = \frac{\exp\left[-\frac{\overline{\gamma}_1+\overline{\gamma}_2-\sqrt{(\overline{\gamma}_1+\overline{\gamma}_2)^2-4\overline{\gamma}_1\overline{\gamma}_2(1-\rho)}}{2\overline{\gamma}_1\overline{\gamma}_2(1-\rho)}\gamma_t\right]-\exp\left[-\frac{\overline{\gamma}_1+\overline{\gamma}_2+\sqrt{(\overline{\gamma}_1+\overline{\gamma}_2)^2-4\overline{\gamma}_1\overline{\gamma}_2(1-\rho)}}{2\overline{\gamma}_1\overline{\gamma}_2(1-\rho)}\gamma_t\right]}{\sqrt{(\overline{\gamma}_1+\overline{\gamma}_2)^2-4\overline{\gamma}_1\overline{\gamma}_2(1-\rho)}}; \gamma_t \geq 0, \quad (12)$$

which itself reduces to the well-known expression for the case of identical Rayleigh channels originally derived in [15] and which can also be found in [3, Eq. (40)]

$$p_a(\gamma_t) = \frac{1}{2\sqrt{\rho}\overline{\gamma}}\left[\exp\left[-\frac{\gamma_t}{(1+\sqrt{\rho})\overline{\gamma}}\right] - \exp\left[-\frac{\gamma_t}{(1-\sqrt{\rho})\overline{\gamma}}\right]\right]; \quad \gamma_t \geq 0. \quad (13)$$

2.1.2. MGF

Substituting (7) in (1) then using the Laplace transform [16, p. 1182, Eq. (110)], it can be shown after some manipulations that the MGF of $p_a(\gamma_t)$ is given by

$$\mathcal{M}_a(s;\overline{\gamma}_1,\overline{\gamma}_2;m;\rho) \triangleq \mathcal{M}_a(s) = \left(1 - \frac{(\overline{\gamma}_1+\overline{\gamma}_2)}{m}\,s + \frac{(1-\rho)\overline{\gamma}_1\overline{\gamma}_2}{m^2}\,s^2\right)^{-m}; \quad s \geq 0. \quad (14)$$

2.2. Model B: D Identically Distributed Branches with Constant Correlation

Model B was proposed by Aalo [6, Section II-A] for identically distributed Nakagami-m channels (i. e., all channels are assumed to have the same average SNR per symbol $\overline{\gamma}$ and the same fading parameter m). This model assumes that the envelope correlation coefficient ρ is the same between all the channel pairs (d,d') $(d,d'=1,2,\cdots,D)$, i.e.

$$\rho = \rho_{dd'} = \frac{\operatorname{cov}(r_d^2, r_{d'}^2)}{\sqrt{\operatorname{var}(r_d^2)\operatorname{var}(r_{d'}^2)}}, \quad d \neq d', \quad 0 \leq \rho < 1, \quad (15)$$

and may therefore correspond to the scenario of multichannel reception from closely placed diversity antennas.

2.2.1. PDF

Based on the work of Gurland [17], Aalo showed the the PDF of γ_t is given in this case by [6, Eq. (18)][2]

$$p_b(\gamma_t) = \frac{\left(\frac{m\gamma_t}{\overline{\gamma}}\right)^{Dm-1}\exp\left(-\frac{m\gamma_t}{(1-\sqrt{\rho})\overline{\gamma}}\right)\,{}_1F_1\left(m, Dm; \frac{Dm\sqrt{\rho}\gamma_t}{(1-\sqrt{\rho})(1-\sqrt{\rho}+D\sqrt{\rho})\overline{\gamma}}\right)}{\left(\frac{\overline{\gamma}}{m}\right)(1-\sqrt{\rho})^{m(D-1)}\,(1-\sqrt{\rho}+D\sqrt{\rho})^m\,\Gamma(Dm)}; \quad \gamma_t \geq 0, \quad (16)$$

where ${}_1F_1(\cdot,\cdot;\cdot)$ is the confluent hypergeometric function [14, Chapter 13, p. 503]. For $D=2$ using [14, Eq. (13.6.3), p. 509], namely

$${}_1F_1(m,2m;2z) = \Gamma\left(m+\frac{1}{2}\right)\left(\frac{z}{2}\right)^{-m+\frac{1}{2}} I_{m-\frac{1}{2}}(z)\,e^z, \quad (17)$$

---

[2] It should be noted at this point that in [6, Eq. (18)] the symbol ρ is used to denote the correlation coefficient of the underlying Gaussian processes that produce the fading on the channels. This correlation coefficient is equal to the square root of the power correlation coefficient. Based on the work of [18, p. 62] it is shown in [3, Appendix V] that for all practical purposes the power correlation coefficient can be assumed to be equal to the envelope correlation coefficient which is denoted by ρ throughout this paper so as to follow what seems to be the more conventional usage of this symbol.

as well as the identities [14, Eq. (6.1.12) and Eq. (6.1.8), p. 255] yielding

$$\Gamma\left(m+\frac{1}{2}\right)=\frac{1\cdot 3\cdot 5\cdot 7\cdots(2m-1)}{2^m}\sqrt{\pi} \tag{18}$$

then it can be shown after some manipulations that (16) reduces to (10) of Model A, as expected.

2.2.2. MGF

Substituting (16) in (1) then using the Laplace transform [16, Eq. (4), p. 864],

$$\int_0^\infty x^{b-1}\ {}_1F_1(a;c;kx)\ e^{-sx}\ dx=\frac{\Gamma(b)}{s^b}\ {}_2F_1\left(a,b;c;\frac{k}{s}\right),\quad |s|>|k|, b>0, s>0, s>k, \tag{19}$$

together with the identity [14, Eq. (15.1.8), p. 556], [6, Eq. (A-5)]

$${}_2F_1(a,b;b;z)=(1-z)^{-a}, \tag{20}$$

it can be shown that

$$\mathcal{M}_b(s;\overline{\gamma};m;\rho;D)\triangleq\mathcal{M}_b(s)=\left(1-\frac{\overline{\gamma}(1-\sqrt{\rho}+D\sqrt{\rho})}{m}\,s\right)^{-m}\left(1-\frac{\overline{\gamma}(1-\sqrt{\rho})}{m}\,s\right)^{-m(D-1)};\quad s\geq 0. \tag{21}$$

For $D=2$, as a check, it can be easily shown that (21) agrees with (14) for $\overline{\gamma}_1=\overline{\gamma}_2$.

2.3. Model C: *D* Identically Distributed Branches with Exponential Correlation

Model C was also proposed by Aalo [6, Section II-B] for identically distributed Nakagami-*m* channels (i. e., all channels are assumed to have the same average SNR per symbol $\overline{\gamma}$ and the same fading parameter m). This model assumes an exponential envelope correlation coefficient $\rho_{dd'}$ between any channel pairs (d,d') $(d,d'=1,2,\cdots,D)$ as given by

$$\rho_{dd'}=\frac{\operatorname{cov}(r_d^2,r_{d'}^2)}{\sqrt{\operatorname{var}(r_d^2)\operatorname{var}(r_{d'}^2)}}=\rho^{|d-d'|}, 0\leq\rho\leq 1, \tag{22}$$

and may therefore correspond to the scenario of multichannel reception from equispaced diversity antennas in which the correlation between the pairs of combined signals decays as the spacing between the antennas increases.

2.3.1. PDF

Based on the work of Kotz and Adams [19], Aalo showed the the PDF of γ_t can be very well approximated by a gamma distribution given by [6, Eq. (19)][3]

$$p_c(\gamma_t)=\frac{\gamma_t^{\frac{mD^2}{r_\rho}-1}\exp\left(-\frac{mD\gamma_t}{r_\rho\overline{\gamma}}\right)}{\Gamma\left(\frac{mD^2}{r_\rho}\right)\left(\frac{r_\rho\overline{\gamma}}{mD}\right)^{\frac{mD^2}{r_\rho}}};\quad \gamma_t\geq 0, \tag{23}$$

where[4]

$$r_\rho=D+\frac{2\sqrt{\rho}}{1-\sqrt{\rho}}\left(D-\frac{1-\rho^{D/2}}{1-\sqrt{\rho}}\right). \tag{24}$$

---

[3] Based on the work of Kotz and Adams [19], Aalo [6] points out that the approximation (23) is valid for high values of D but is still accurate for values of D as small as 5.

[4] We remind the reader that contrary to its usage in [6] the coefficient ρ denotes in this paper the envelope correlation coefficient.

2.3.2. *MGF*

Substituting (23) in (1) then using the Laplace transform [16, Eq. (3.381.4)], namely

$$\int_0^\infty x^{\nu-1}\, e^{-sx}\, dx = \frac{\Gamma(\nu)}{s^\nu}, \tag{25}$$

it can be shown that

$$\mathcal{M}_c(s;\overline{\gamma};m;\rho;D) \triangleq \mathcal{M}_c(s) = \left(1 - \frac{r_\rho \overline{\gamma}}{mD}\, s\right)^{-\frac{mD^2}{r_\rho}}; \quad s \geq 0. \tag{26}$$

2.4. Model D: D Nonidentically Distributed Branches with Arbitrary Correlation

In this section we treat a very general model in which the combined branches may not be identically distributed and may have an arbitrary correlation. More specifically this model assumes that the branches have an average SNR per symbol $\overline{\gamma}_d$ and the same fading parameter m. The envelope correlation coefficient $\rho_{dd'}$ between any channel pairs (d, d') $(d, d' = 1, 2, \cdots, D)$ is denoted by $\rho_{dd'}$. The PDF of the combined SNR corresponding to this model was not found in a simple form. However the joint distribution of the $\{\gamma_d\}_{d=1}^D$ can be deduced from the work of Krishnamoorthy and Parthasarathy [20] and can be expressed in terms of the generalized Laguerre polynomials. Their derivation is based on a relatively simple form for the joint MGF of a D-variate gamma distribution. Based on that derivation and defining $\beta_d = \frac{m\gamma_d}{\overline{\gamma}_d}$ as the normalized SNR per symbol and per branch we can express the MGF corresponding to this model as

$$\mathcal{M}_d(s;\{\overline{\gamma}_d\}_{d=1}^D;m;[\rho_{dd'}];D) \triangleq \mathcal{M}_d(s) = E_{\gamma_1,\gamma_2,\cdots,\gamma_D}\left[\exp\left(s\sum_{d=1}^D \gamma_d\right)\right] = E_{\beta_1,\beta_2,\cdots,\beta_D}\left[\exp\left(s\sum_{d=1}^D \frac{\overline{\gamma}_d}{m}\beta_d\right)\right]. \tag{27}$$

Using the result [20, Eq. (2.3)] we can rewrite (27) as

$$\mathcal{M}_d(s) = \prod_{d=1}^D \left(1 - \frac{s\overline{\gamma}_d}{m}\right)^{-m} \left|\begin{bmatrix} 1 & \sqrt{\rho_{12}}\left(1-\frac{m}{s\overline{\gamma}_2}\right)^{-1} & \cdots & \sqrt{\rho_{1D}}\left(1-\frac{m}{s\overline{\gamma}_D}\right)^{-1} \\ \sqrt{\rho_{12}}\left(1-\frac{m}{s\overline{\gamma}_1}\right)^{-1} & 1 & \cdots & \sqrt{\rho_{2D}}\left(1-\frac{m}{s\overline{\gamma}_D}\right)^{-1} \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{\rho_{1D}}\left(1-\frac{m}{s\overline{\gamma}_1}\right)^{-1} & \sqrt{\rho_{2D}}\left(1-\frac{m}{s\overline{\gamma}_2}\right)^{-1} & \cdots & 1 \end{bmatrix}\right|_{D\times D}^{-m}, \tag{28}$$

where $|[M]|_{D\times D}$ denotes the determinant of the $D \times D$ matrix M. For $D = 2$, as a check, it is straightforward to show that (28) reduces to (14) of model A. For the case of identical channels $(\overline{\gamma}_d = \overline{\gamma})$, (28) reduces after some manipulations to

$$\mathcal{M}_d(s) = \left(-\frac{s\overline{\gamma}}{m}\right)^{-mD} \left|\begin{bmatrix} 1-\frac{m}{s\overline{\gamma}} & \sqrt{\rho_{12}} & \cdots & \sqrt{\rho_{1D}} \\ \sqrt{\rho_{12}} & 1-\frac{m}{s\overline{\gamma}} & \cdots & \sqrt{\rho_{2D}} \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{\rho_{1D}} & \sqrt{\rho_{2D}} & \cdots & 1-\frac{m}{s\overline{\gamma}} \end{bmatrix}\right|_{D\times D}^{-m}. \tag{29}$$

3. COHERENT MULTICHANNEL RECEPTION

To assess the average error probability performance of a diversity receiver in the presence of slow fading we must first provide an expression for the conditional (on γ_t) error probability and then average this expression over the PDFs of γ_t as given in Section 2, i.e.,

$$P_i(E) = \int_0^\infty P(E|\gamma_t)\, p_i(\gamma_t)\, d\gamma_t. \tag{30}$$

We now present specific results for different coherent modulations with MRC reception.

3.1. Binary Phase-Shift-Keying (BPSK) and Binary Frequency-Shift-Keying (BFSK)

Based on an alternate representation of the Gaussian Q-function [21], the conditional BER of BPSK and BFSK can be written in the integral form [1]

$$P_b(E|\gamma_t) = \frac{1}{\pi}\int_0^{\pi/2} \exp\left(-\frac{g\gamma_t}{\sin^2\phi}\right) d\phi, \tag{31}$$

where $g = 1$ for BPSK, $g = 1/2$ for orthogonal BFSK, and $g = \frac{1}{2} + \frac{1}{3\pi}$ for BFSK with minimum correlation. The average BER, $P_b(E)$ is then obtained by averaging (31) over (7), (16), or (23) yielding after interchanging the order integration the desired generic result

$$P_{b_i}(E) = \frac{1}{\pi}\int_0^{\pi/2} \mathcal{M}_i\left(-\frac{g}{\sin^2\phi}\right) d\phi. \tag{32}$$

For model B, (32) with (21) should be contrasted with the previous available equivalent result [6, Eq. (32)] which required the evaluation of the Appell's hypergeometric function, $F_2(\cdot;\cdot,\cdot;\cdot,\cdot)$, which is typically not available in standard mathematical software libraries such us Mathematica, Matlab, and Maple, and which is defined either in term of an infinite range integral of a special function [6, Eq. (A-12)] or alternatively as a double infinite sum [6, Eq. (A-13)]. For model C, (32) along with (26) is equivalent to [6, Eq. (40)] which is expressed in terms of the hypergeometric function ${}_2F_1(\cdot,\cdot;\cdot)$ [14, Chapter 15, p. 555].

3.2. Multiple Phase-Shift-Keying (M-PSK)

The generalization of (31) to M-ary signaling is straightforward in view of the form of the symbol error rate for M-PSK on the AWGN [22, Eq. (71)], which on the fading channel becomes the conditional SER [1]

$$P_s(E|\gamma_t) = \frac{1}{\pi}\int_0^{(M-1)\pi/M} \exp\left(-\frac{g_{\text{psk}}\gamma_t}{\sin^2\phi}\right) d\phi, \tag{33}$$

where $g_{\text{psk}} = \sin^2\left(\frac{\pi}{M}\right)$ Thus, by inspection we immediately obtain

$$P_{s_i}(E) = \frac{1}{\pi}\int_0^{(M-1)\pi/M} \mathcal{M}_i\left(-\frac{g_{\text{psk}}}{\sin^2\phi}\right) d\phi. \tag{34}$$

It should be noted that for model A (34) together with (14) is an *exact* expression equivalent to [4, Eq. (54)] with the advantage of being expressed in terms of a single finite-range integral (contrary to [4, Eq. (54)]) and whose integrand is much simpler than [4, Eq. (54)], and hence easier to compute for any arbitrary value of the fading parameter m.

3.3. *M*-ary Quadrature Amplitude Modulation (*M*-QAM)

Consider square M-QAM signals whose constellation size is given by $M = 2^k$ with k even. For these type of constellations, the expression for the SER on the AWGN involves both the Gaussian Q-function and the square of this function. Using an alternate representation for the square of the Gaussian Q-function [23, Eq. (80)], the conditional SER for the fading channel is given by [1]

$$P_s(E|\gamma t) = \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2} \exp\left(-\frac{g_{\text{qam}}\gamma_t}{\sin^2\phi}\right) d\phi - \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2 \int_0^{\pi/4} \exp\left(-\frac{g_{\text{qam}}\gamma_t}{\sin^2\phi}\right) d\phi, \quad (35)$$

where $g_{\text{qam}} = 3/\left[2(M-1)\right]$. Hence, by inspection, the average SER is given by

$$P_{s_i}(E) = \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2} \mathcal{M}_i\left(-\frac{g_{\text{qam}}}{\sin^2\phi}\right) d\phi - \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2 \int_0^{\pi/4} \mathcal{M}_i\left(-\frac{g_{\text{qam}}}{\sin^2\phi}\right) d\phi. \quad (36)$$

4. DIFFERENTIALLY COHERENT AND NONCOHERENT MULTICHANNEL RECEPTION

For differentially coherent and noncoherent detection, MRC is not practical since it requires channel phase estimates for its implementation, and post-detection EGC (with differential or noncoherent square-law detection) is in this case the more common form of diversity combining. We present in this section a generic result for the average BER for differentially coherent and noncoherent multichannel receivers when used in conjunction with post-detection EGC over correlated Nakagami-m fading channels.

4.1. Conditional BER

Using an alternate integral representation of the generalized Marcum Q-function [24, Appendix C], [25], the authors showed that the conditional BER of a wide class of differentially coherent and noncoherent modulations can be written in the desired form consisting of a single integral with finite limits and an integrand which is composed of elementary functions and where the total instantaneous SNR per bit γ_t (over which we must average) appears only in the argument of the exponential term, namely,

$$P_b(E|\gamma_t) = \frac{b^2\eta^L}{2\pi(1+\eta)^{2L-1}} \int_{-\pi}^{\pi} \frac{f\left(L; \frac{a}{b}, \eta; \phi\right)}{(a^2+b^2+2ab\sin\phi)} \exp\left[-(a^2+b^2+2ab\sin\phi)\frac{\gamma_t}{2}\right] d\phi, 0^+ \le a < b, \quad (37)$$

where

$$f(L;\beta,\eta;\phi) = f_0(L;\beta,\eta;\phi) + f_1(L;\beta,\eta;\phi),$$

with

$$f_0(L;\beta,\eta;\phi) = \left[-\frac{(1+\eta)^{2L-1}}{\eta^L} + \sum_{l=1}^{L}\binom{2L-1}{L-l}(\eta^{-l}+\eta^{l-1})\right] \beta(\beta+\sin\phi),$$

$$f_1(L;\beta,\eta;\phi) = \sum_{l=1}^{L}\binom{2L-1}{L-l}\left[(\eta^{-l}\beta^{-l+1}-\eta^{l-1}\beta^{l+1})\cos((l-1)(\phi+\pi/2)) - (\eta^{-l}\beta^{-l+2}-\eta^{l-1}\beta^{l})\cos(l(\phi+\pi/2))\right].$$

In (37) the parameters a and b are modulation-dependent and are defined in [12, Eq. (4B.22)], and $\eta = v_2/v_1$, with the parameters v_1, v_2 defined in A number of special cases are of particular importance. For noncoherent detection of equal energy, equiprobable, correlated binary signals, $\eta = 1$ and

$$a = \left(\frac{1-\sqrt{1-|\lambda|^2}}{2}\right)^{1/2} \quad b = \left(\frac{1+\sqrt{1-|\lambda|^2}}{2}\right)^{1/2}, \quad (38)$$

where λ $(0 \leq |\lambda| \leq 1)$ is the complex-valued cross-correlation coefficient between the two signals. The special case $\lambda = 0$ corresponds to orthogonal noncoherent BFSK for which $a = 0$ and $b = 1$. Furthermore, in the case of binary differential phase-shift-keying (BDPSK), $a = 0$, $b = \sqrt{2}$, and $\eta = 1$. Finally $a = \sqrt{2 - \sqrt{2}}$, $b = \sqrt{2 + \sqrt{2}}$, and $\eta = 1$ correspond to quaternary differential phase-shift-keying (DQPSK) with Gray coding. It should be noted that as $a \to 0$ (37) assumes an indeterminate form but the limit converges smoothly to the exact conditional BER expression.

4.2. Average BER

To evaluate the average BER one must average the conditional BER expression (37) over one of the fading PDFs under consideration as given by (7), (16), or (23). Recognizing the analogy between (37) and (31) in so far as its functional dependence on γ_t, we can immediately write the average BER as

$$P_{b_i}(E) = \frac{b^2 \eta^L}{2\pi(1+\eta)^{2L-1}} \int_{-\pi}^{\pi} \frac{f\left(L; \frac{a}{b}, \eta; \phi\right)}{(a^2 + b^2 + 2ab\sin\phi)} \mathcal{M}_i\left(-\frac{a^2 + b^2 + 2ab\sin\phi}{2}\right) d\phi. \qquad (39)$$

Letting $a \to 0$ and $b = 1$ $(b = \sqrt{2})$ in (39) yields the average BER performance of orthogonal BFSK (BDPSK), and in this particular case (39) is equivalent to the closed-forms [7, Eq. (21)] and [7, Eq. (16)], for models A and B, respectively.

5. *REFERENCES*

[1] M. -S. Alouini and A. Goldsmith, "A unified approach for calculating the error rates of linearly modulated signals over generalized fading channels," in *Proc. IEEE Int. Conf. on Commun. ICC'98, Atlanta, GA*, pp. 459–464, June 1998.

[2] M. K. Simon and M. -S. Alouini, "A unified approach for the probability of error for noncoherent and differentially coherent modulations over generalized fading channels." To appear in *IEEE Trans. Commun.*

[3] J. N. Pierce and S. Stein, "Multiple diversity with nonindependent fading," *Proc. IRE*, vol. 48, pp. 89–104, January 1960.

[4] Y. Miyagaki, N. Morinaga, and T. Namekawa, "Error probability characteristics for CPSK signal through m-distributed fading channel," *IEEE Trans. Commun.*, vol. COM-26, pp. 88–100, January 1978.

[5] E. K. Al-Hussaini and A. M. Al-Bassiouni, "Performance of MRC diversity systems for the detection of signals with Nakagami fading," *IEEE Trans. Commun.*, vol. COM-33, pp. 1315–1319, December 1985.

[6] V. A. Aalo, "Performance of maximal-ratio diversity systems in a correlated Nakagami-fading environment," *IEEE Trans. Commun.*, vol. COM-43, pp. 2360–2369, August 1995.

[7] F. Patenaude, J. H. Lodge, and J. -Y. Chouinard, "Error probability expressions for non-coherent diversity in Nakagami fading channels," in *Proc. IEEE Veh. Technol. Conf. VTC'97, Phoenix, AZ*, pp. 1484–1487, May 1997.

[8] F. Patenaude, J. H. Lodge, and J. -Y. Chouinard, "Noncoherent diversity reception over Nakagami-fading channels," *IEEE Trans. on Commun.*, vol. COM-46, pp. 985–991, August 1998.

[9] M. Nakagami, "The m-distribution- A general formula of intensity distribution of rapid fading," in *Statistical Methods in Radio Wave Propagation*, pp. 3–36, Pergamon Press, Oxford, U. K., 1960.

[10] D. Brennan, "Linear diversity combining techniques," *Proc. IRE*, vol. 47, pp. 1075–1102, June 1959.

[11] G. L. Stüber, *Principles of Mobile Communications.* Norwell, MA: Kluwer Academic Publishers, 1996.

[12] J. G. Proakis, *Digital Communications.* New York, NY: McGraw-Hill, second ed., 1989.

[13] M. -S. Alouini and M. K. Simon, "Multiple diversity over correlated Nakagami-m fading channels." In preparation for submission to *IEEE Transactions.*

[14] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York, NY: Dover Publications, ninth ed., 1970.

[15] K. S. Packard, "Effect of correlation on combiner diversity," *Proc. IRE*, vol. 46, pp. 362–363, January 1958.

[16] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products.* San Diego, CA: Academic Press, fifth ed., 1994.

[17] J. Gurland, "Distribution of the maximum of the arithmetic mean of correlated random variables," *Annals of Math. Statist.*, vol. 26, pp. 294–300, 1955.

[18] J. A. Lawson and G. E. Uhlenbeck, *Threshold Signals.* New York, NY: McGraw-Hill Book Co., 1952.

[19] S. Kotz and J. Adams, "Distribution of sum of identically distributed exponentially correlated gamma variables," *Annals of Math. Statist.*, vol. 35, pp. 227–283, June 1964.

[20] A. S. Krishnamoorthy and M. Parthasarathy, "A multivariate gamma-type distribution," *Annals Math Statist.*, vol. 22, pp. 549–557, 1951.

[21] J. W. Craig, "A new, simple, and exact result for calculating the probability of error for two-dimensional signal constellations," in *Proc. IEEE Milit. Commun. Conf. MILCOM'91, McLean, VA*, pp. 571–575, October 1991.

[22] R. F. Pawula, S. O. Rice, and J. H. Roberts, "Distribution of the phase angle between two vectors perturbed by Gaussian noise," *IEEE Trans. Commun.*, vol. COM-30, pp. 1828–1841, August 1982.

[23] M. K. Simon and D. Divsalar, "Some new twists to problems involving the Gaussian probability integral," *IEEE Trans. Commun.*, vol. COM-46, pp. 200–210, February 1998.

[24] C. W. Helstrom, *Elements of Signal Detection and Estimation.* Englewood Cliffs, NJ: PTR Prentice-Hall, 1995.

[25] M. K. Simon, "A new twist on the Marcum Q-function and its applications," *IEEE Commun. Letters*, vol. 2, pp. 39–41, February 1998.

Improved Performance of a PSAM System with Iterative Filtering and Decoding

Hsuan-Jung Su and Evaggelos Geraniotis
Dept. of Electrical Engineering and Institute for Systems Research
University of Maryland at College Park
College Park, MD 20742
hjsu@eng.umd.edu, evagelos@eng.umd.edu

Abstract

Since its first appearance, turbo code has repeatedly proved its superiority over other coding schemes. The key to its good performance lies in the passing and feedback of the soft information, which allow the advantages of concatenation to be fully explored. In time-varying channels such as fading channels, the exact channel status is no longer constant. As noted in previous works, channel side information (CSI) becomes crucial for obtaining reliable soft information and good decoding performance. In this paper, we attempt to perform CSI estimation together with the decoding process so both parts can benefit each other. The idea is akin to the iterative approach of joint equalization and decoding. For channel probing, we apply pilot symbol assisted modulation (PSAM) for it has been used successfully in providing coherent demodulation under fading channels. In view of the time-varying nature of the wireless mobile channel statistics, the PSAM filters in our design are made adaptive. Moreover, they are designed to be able to exploit the soft information generated by the iterative decoding. With this joint design of iterative filtering and decoding, the decoding performance is shown to be better than conventional turbo coded PSAM systems. The performance gap between practical systems and the hypothetical system with perfect CSI is therefore made smaller. More importantly, this design has linear complexity and does not alter much the existing PSAM and iterative decoding schemes. Considering that the optimal joint design of channel estimation and decoding has exponential complexity, the proposed system provides a simple alternative while maintaining an effective performance improvement.

1 Introduction

To date, the excellence of turbo codes has been recognized and agreed in the communication society. Most of the works done, however, were under additive white Gaussian noise (AWGN) channels where the channel statistics are time-invariant. Only recently, researchers started investigating the performance of turbo codes under fading environments. It was shown that the mismatch in CSI will worsen the decoding performance[1, 2]. Therefore, obtaining correct CSI becomes vital in designing a turbo coded system.

Among those previous studies, some assumed perfect CSI; others had imperfect CSI estimation but coherent demodulation. These assumptions seem idealistic for practical systems. In the typical operating range of turbo codes, the signal to noise ratio (SNR) is usually too low to have perfect channel estimates. When we consider the reverse link of CDMA systems, the situation becomes even worse. The signal strength before despreading does not even guarantee accurate carrier phase tracking. In view of the above problems, we propose the inclusion of PSAM for coherent demodulation and channel estimation. The idea of PSAM is very similar to the pilot signal utilized in the forward link communication where it poses as channel reference. It was shown in [3] that good performance can be achieved with reasonable throughput loss. In moderate Doppler conditions, PSAM outperforms differential modulation by avoiding error propagation.

Despite its advantages, some weaknesses are still being observed in the PSAM design. Namely, its lack of adaptability to the time-varying channel statistics and large throughput loss when the Doppler bandwidth is large. The first weakness comes from the fact that the channel estimation filter is optimized at a pre-defined noise and Doppler condition. As pointed out in [3], if the Doppler bandwidth is beyond the operating point of the filter, the performance is highly deteriorated. Therefore, a "worst case" filter design is desired, which means more loss of throughput, and, the filter is always over estimating the Doppler bandwidth. The second weakness is inherent in PSAM systems. It is this shortage which limits the application of PSAM to only slow fading channels.

When we combine PSAM with iterative decoding, the two designs can benefit each other. For the iterative (soft) decoding, the noise, channel phase and amplitude information are supplied by the PSAM filtering. For the PSAM filters, the iterative decoding basically provides the reliability of the filter outputs which can help improving the filter tap weights. The time required for the decoding iterations is not "wasted" because it is also the time for the filter adaptation to converge. Furthermore, after the first decoding iteration, the PSAM filter can use all received symbols as references since it has obtained all their reliability information from the decoder. The PSAM sampling rate is essentially boosted, so the operating range of the PSAM design is widened.

The idea of concatenating two estimation schemes is not uncommon in turbo coded system designs[4, 5]. What prevents it from being practical is its exponentially growing complexity. For this reason, one of the two features of this paper aims at designing a suboptimal realization with linear complexity. The other feature, namely, filter adaptability, comes naturally from the linearized structure.

2 System Model

In this section, we first recapitulate the commonly used fading channel model and the notations. Then the previously proposed PSAM design and the issues of importance in turbo decoding are briefly described. We demonstrate our design of combined adaptive filtering and iterative decoding in the next section. Its complexity and some performance aspects will be discussed.

2.1 Time Correlated Fading Channel

Let $s(t)$ be the input signal to the fading channel, and $r_c(t)$ denote the fading channel output signal given by

$$r_c(t) = c(t)s(t) + n_c(t), \tag{1}$$

where $n_c(t)$ is the AWGN with two-sided spectral density $N_0/2$ in both real and imaginary components. The channel gain $c(t)$ consists of both amplitude gain and phase shift, therefore it is also complex. In this paper, we consider the channel without line of sight (LOS), which is usually the case for wireless communication in urban area. The case with LOS (i.e., Rician channel) is just a direct extension. For simplicity, we consider flat fading only, which means the channel symbol (in CDMA, it is chip) duration is larger than the excess delay of the multi-path power delay profile. For cases when multi-path is resolvable, some diversity combing skills might be used. The essence of our design, however, does not change since the filters are adaptive.

The complex channel gain $c(t)$ incorporates both fading and frequency offset:

$$c(t) = \exp(j2\pi f_0 t)g(t) \tag{2}$$

where f_0 is the residual frequency offset, and $g(t)$ is the complex Gaussian fading process with variance σ_g^2 and Doppler spread f_D. Its autocorrelation function can be written as

$$R_c(\tau) = \sigma_g^2 \tilde{R}_c(\tau) \tag{3}$$

where $\tilde{R}_c(\tau)$ is the unit power equivalence. For Rayleigh fading channel, according to [6],

$$\tilde{R}_c(\tau) = \exp(j2\pi f_0 \tau) J_0(2\pi f_D \tau) \tag{4}$$

where $J_0(\cdot)$ is the zero-th order Bessel function of the first kind.

2.2 Pilot Symbol Assisted Modulation and Channel Estimation

In PSAM, redundant symbols with known values are embedded to the transmission signal once per $M-1$ data symbols. At the receiver, these known symbols are filtered (interpolated) to estimate the channel gain of the data symbols. Coherent demodulation is therefore enabled by rotating the received signal phase back through use of the channel estimates.

Since the channel gain and the received signal are complex Gaussian distributed, Wiener filter is optimum. Consider that a Wiener filter with K taps is used to estimate the channel gain of the kth data symbol in a frame of M symbols. Let the length K column vector $\mathbf{r}$ be the set of nearest pilot symbols (with known value "+1") to the data symbol of interest. Then the estimated channel gain of the kth data symbol is given by

$$v(k) = \mathbf{h}^H(k)\mathbf{r} \tag{5}$$

where $\mathbf{h}$ is the vector of Wiener filter coefficients obtained by solving

$$\mathbf{R}\mathbf{h}(k) = \mathbf{w}(k). \tag{6}$$

In (6), $\mathbf{R}$ is the $K \times K$ autocorrelation matrix defined as

$$\mathbf{R} = E\left[\mathbf{r}\mathbf{r}^H\right] \tag{7}$$

and $\mathbf{w}(k)$ is a length K covariance vector given by

$$\mathbf{w}(k) = E\left[\mathbf{r}c^*(k)\right] \tag{8}$$

where $c(k)$ is the desired channel gain. $\mathbf{R}$ and $\mathbf{w}(k)$ can be computed through (1)-(4) given that the Doppler bandwidth f_D, the frequency offset f_0, and the AWGN power are known a priori. Once $\mathbf{R}$ and $\mathbf{w}(k)$ are ready, we can compute the optimal Wiener filter $\mathbf{h}$ and apply it to obtain the channel gain estimate $v(k)$. The decision is then made based on the corrected signal $r(k)v^*(k)$.

2.3 Turbo Decoding in Time Correlated Fading Channels

In turbo decoding, we are interested in getting the log-likelihood ratio (LLR) of the information bits given the received signal. Let $\mathbf{y}$ be the vector of the received sequence and u_k be an information bit, the LLR associated with u_k is computed as

$$L(\hat{u}_k) = L(u_k) + L(\tilde{x}_{k;1}) + L_e(\hat{u}_k) \tag{9}$$

where $L(u_k)$ is the LLR of the a priori probability of u_k usually supplied by another constituent decoder; $L(\tilde{x}_{k;\nu})$ is given by

$$L(\tilde{x}_{k;\nu}) = \log \frac{P(y_{k;\nu}|x_{k;\nu} = +1)}{P(y_{k;\nu}|x_{k;\nu} = -1)} \tag{10}$$

in which $x_{k;\nu}, 1 \leq \nu \leq N$, are the code bits associated with u_k; and $y_{k;\nu}$ are their received (complex) values after the channel distortion. $L_e(\hat{u}_k)$ is the extrinsic information obtained through the code constraint. From (9) and (10) it is clear that we need good channel estimation in order to compute $L(\tilde{x}_{k;\nu})$ and hence $L(\hat{u}_k)$. For BPSK, given that the channel gain for $x_{k;\nu}$ is $c_{k;\nu}$ and the AWGN variance per dimension is σ_n^2, (10) can be written as

$$L(\tilde{x}_{k;\nu}) = \frac{2Re\left[y_{k;\nu}c^*_{k;\nu}\right]}{\sigma_n^2}. \tag{11}$$

Unfortunately, the channel gain is time correlated, and all we have is the noisy observation $\mathbf{y}$. For an optimum scheme, we would need to compute

$$L(\tilde{x}_{k;\nu}) = \log \frac{\sum_{\mathbf{x}'} P(y_{k;\nu}, \mathbf{x}', \mathbf{y}'|x_{k;\nu} = +1)}{\sum_{\mathbf{x}'} P(y_{k;\nu}, \mathbf{x}', \mathbf{y}'|x_{k;\nu} = -1)} \tag{12}$$

where $\mathbf{x}'$ and $\mathbf{y}'$ are the code sequence and the received sequence respectively with the bit in question (i.e., $x_{k;\nu}, y_{k;\nu}$) excluded; and the expression for the extrinsic information $L_e(\hat{u}_k)$ would need to be modified[5]. This scheme has exponential complexity, which prevents us from using the entire received sequence for the estimation. It was said in [5] that using more than 5 consecutive symbols for the estimation will result in an inordinate amount of computation. Truncating the number of reference symbols worsens the channel estimation accuracy. When the fading is slow and the channel time correlation is high, the performance of the truncated scheme gets even farther behind that of the optimum scheme.

3 Coded PSAM System with Iterative Filtering and Decoding

In this section, we describe our design of a linear complexity adaptive channel estimator which works jointly with the iterative decoder while maintaining least modification of the standard schemes. Before getting into the details, we would like to reiterate our motivations again. The goal is to let PSAM and iterative decoding benefit from each other while avoiding high complexity and major modifications in both schemes. For the decoding, we thus provide the crucial CSI. For the PSAM channel estimation, iterative decoding establishes a handy structure for adaptive filtering. As the iterations continue, the PSAM filters can use all received symbols as references. The PSAM sampling rate is thus highly increased, so is the estimation performance.

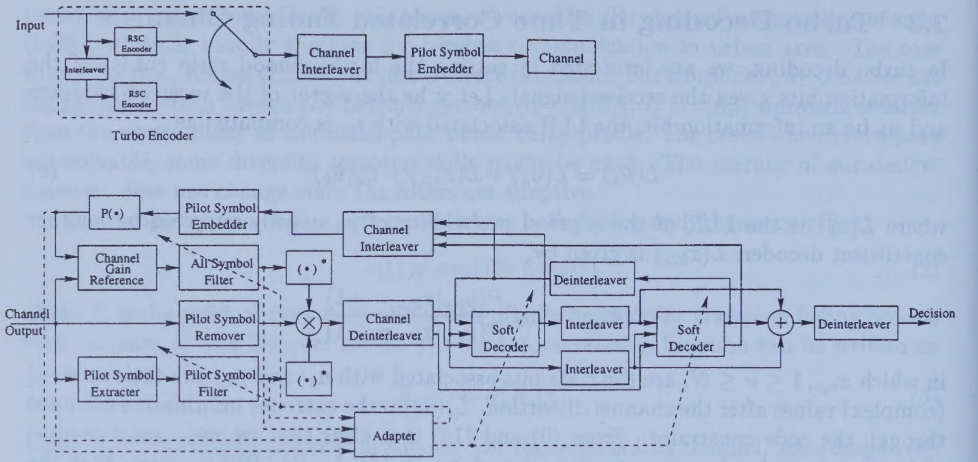

Input
RSC Encoder
Interleaver
RSC Encoder
Turbo Encoder
Channel Interleaver
Pilot Symbol Embedder
Channel
P(•)
Pilot Symbol Embedder
Channel Interleaver
Channel Gain Reference
All Symbol Filter
(•)*
Deinterleaver
Channel Output
Pilot Symbol Remover
Channel Deinterleaver
Soft Decoder
Interleaver
Soft Decoder
Deinterleaver
Decision
Interleaver
Pilot Symbol Extracter
Pilot Symbol Filter
(•)*
Adapter

Figure 1: Turbo Coded PSAM System with Iterative Filtering and Decoding

The proposed system is depicted in Fig. 1, in which the receiver consists of a filter for pilot symbols, a filter for all (pilot and data) symbols, a filter taps updater, a noise power estimator, and a modified turbo decoder. Each time a block of symbols comes in, the pilot symbol filter (PSF) is used as in standard PSAM systems to get the channel gain estimates for the data symbols. These estimates are then deinterleaved with the channel interleaver[7] and passed to the turbo decoder as CSI along with the deinterleaved data sequence. In each decoding iteration, a standard turbo decoder would generate the extrinsic information for the information bits which are to be used in the next iteration as the a priori information. In our design, we need also the extrinsic information for the code bits. The modifications for both Maximum A Posteriori (MAP) and Soft Output Viterbi Algorithm (SOVA) decoders[8, 9] are pretty simple. For example, in a MAP decoder, we have for the information bits

$$L_e(\hat{u}_k) = \log \frac{\sum_{\substack{(s',s)\\ u_k=+1}} \gamma_k^e(s',s)\cdot\alpha_{k-1}(s')\cdot\beta_k(s)}{\sum_{\substack{(s',s)\\ u_k=-1}} \gamma_k^e(s',s)\cdot\alpha_{k-1}(s')\cdot\beta_k(s)} \tag{13}$$

where

$$\gamma_k^e(s',s) = \exp\left(\sum_{\nu=2}^{N} \frac{1}{2}\cdot L(\tilde{x}_{k;\nu})\cdot x_{k;\nu}\right). \tag{14}$$

To get the extrinsic information of the code bits, we need to add the following computations

$$L_e(\hat{x}_{k;\nu}) = \log \frac{\sum_{\substack{(s',s)\\ x_{k;\nu}=+1}} \gamma_k^{e_c}(s',s)\cdot\alpha_{k-1}(s')\cdot\beta_k(s)}{\sum_{\substack{(s',s)\\ x_{k;\nu}=-1}} \gamma_k^{e_c}(s',s)\cdot\alpha_{k-1}(s')\cdot\beta_k(s)} \tag{15}$$

with

$$\gamma_k^{e_c}(s',s) = \exp\left(\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{N} \frac{1}{2}\cdot L(\tilde{x}_{k;\mu})\cdot x_{k;\mu} + \frac{1}{2}\cdot L(u_k)\cdot u_k\right). \tag{16}$$

The extrinsic information serve as the a priori information in the next filtering and decoding iteration. Starting from the second iteration, we use the all symbol filter (ASF) instead of the PSF. The ASF takes all symbols adjacent to the symbol of interest as references to estimate the channel gain for that symbol. The channel gains for the reference symbols can be obtained by rotating the received symbols according to the a priori information. For simplicity, we consider only BPSK in this paper; the essence is fully covered, however. Let p be the a priori probability that a "+1" was transmitted. As in the iterative decoding, we can substitute it with the extrinsic information of the previous decoding iteration. Thus we have

$$p^d \equiv P(x_{k;\nu} = +1) = \frac{1}{1 + \exp\left(-L_e(\hat{x}_{k;\nu})\right)} \tag{17}$$

for the data symbols; and

$$p^p = 1 \tag{18}$$

for the pilot symbols. These a priori probabilities are re-interleaved with the channel interleaver to attain the original order as the channel output sequence. The channel gain (corrupted by AWGN) of a reference symbol is then approximated as

$$c^r(l) = p \cdot y(l) - (1 - p) \cdot y(l). \tag{19}$$

To get the channel gain estimation of a specific symbol, we use the ASF on the reference channel gains

$$v(k) = \mathbf{h}_a^H \mathbf{r}_a(k) \tag{20}$$

where $\mathbf{r}_a(k) = [c^r(k - \frac{K_a}{2}), \ldots, c^r(k-1), c^r(k+1), \ldots, c^r(k + \frac{K_a}{2})]^T$ and $\mathbf{h}_a$ is the ASF with filter length K_a. Note that $\mathbf{h}_a$ was initially optimized with procedures similar to (6)-(8).

Eq. (19) although changes the AWGN distribution, it has the advantage of decreasing the contributions of the unreliable symbols. When the decoding reliability becomes high enough after some iterations, (19) is, in effect, no longer an approximation. The new channel LLR $L(\tilde{x}_{k;\nu})$ to be used in the following decoding are computed by substituting $c_{k;\nu}$ with $v_{k;\nu}$ in (11). After the decoding, new extrinsic information and hence the a priori probabilities are generated. The ASF is adapted with steepest decent method[10], that is,

$$\mathbf{h}_{a;n+1} = \mathbf{h}_{a;n} + \mu_a \cdot E\left[\mathbf{r}_{a;n} e_n^*\right] \tag{21}$$

where n is the index of iteration; μ_a is the step-size parameter; and e_n is the estimation error given by

$$e_n(k) = p(k)\left(y(k) - v(k)\right) + \left(1 - p(k)\right)\left(-y(k) - v(k)\right). \tag{22}$$

The expectation in (21) can be obtained by averaging over the entire code block. The noise power estimate is given by

$$\sigma_{ae}^2 = E\left[e_n e_n^*\right] \tag{23}$$

which again, can be obtained by averaging. The updated ASF is applied to the computation of the new channel gain estimates (20). These new channel estimates, together with the noise estimate (23) are used to generate the new channel LLR $L(\tilde{x}_{k;\nu})$ for the upcoming decoding.

In addition to the adaptation of the ASF, we need to adapt the PSF, too, since we will need initial channel estimates in decoding the next block of sequence. This is done with the same procedures as (17)-(23) after the entire decoding finishes. Except for now the corresponding filter, tap inputs, step-size parameter, and noise power estimate are for the PSF only, they are denoted by $\mathbf{h}_p$, $\mathbf{r}_p$, μ_p, and σ_{pe}^2, respectively.

The two updated filters $\mathbf{h}_a$ and $\mathbf{h}_p$, and the two updated noise estimates σ_{ae}^2 and σ_{pe}^2, are all being stored and carried over to decode the next incoming block. We call the PSF adaptation "inter-block" adaptation in contrast to the "intra-block" ASF adaptation mentioned above. The reason why we use different noise estimates in the first round decoding iteration (σ_{pe}^2) and the remaining iterations (σ_{ae}^2) is to improve the performance, since the initial channel estimation by the PSF is usually much worse.

4 Simulation Parameters and Numerical Results

In order to understand the behavior of the proposed design better, we simulated both convolutional coded and turbo coded PSAM systems. Same systematic encoder was used in both systems; the code rates and the code block sizes of them were also set to be the same. To be specific, the convolutional code was $(1, 5_8/7_8)$, rate 1/2 with block size 1028 bits. Tailbitting was implemented to simplify the decoder. The turbo code was $(1, 5_8/7_8, 5_8/7_8)$, rate 1/2. In other words, the non-systematic parts of the two constituent codes (CC) were punctured every other bit. A random interleaver with block size 1028 was used for the turbo code.

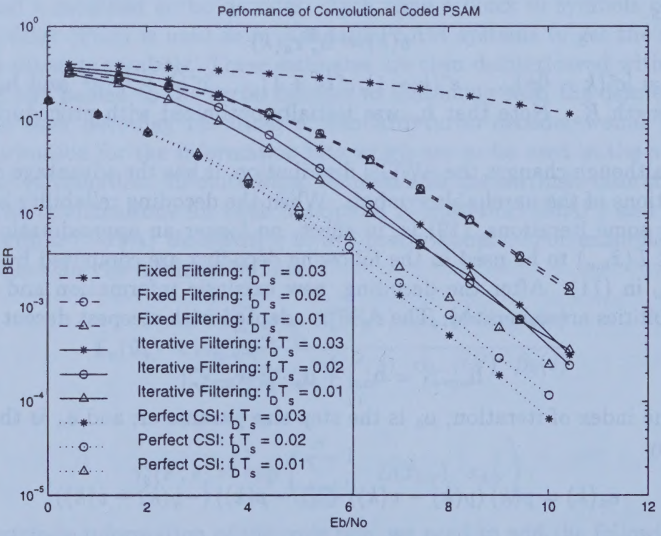


Figure 2: Performance of convolutional coded PSAM systems with one PSF for all symbol positions and an additional ASF for the iterative schemes. The Filters are initially optimized for $f_D T_s = 0.02$ and $E_b/N_0 = 6$ dB

In order to break the channel time correlation, the coded sequence was interleaved before transmission by a 33 × 64 block interleaver, which has a small portion unused. After the channel interleaving, pilot symbols with know values were embedded with a

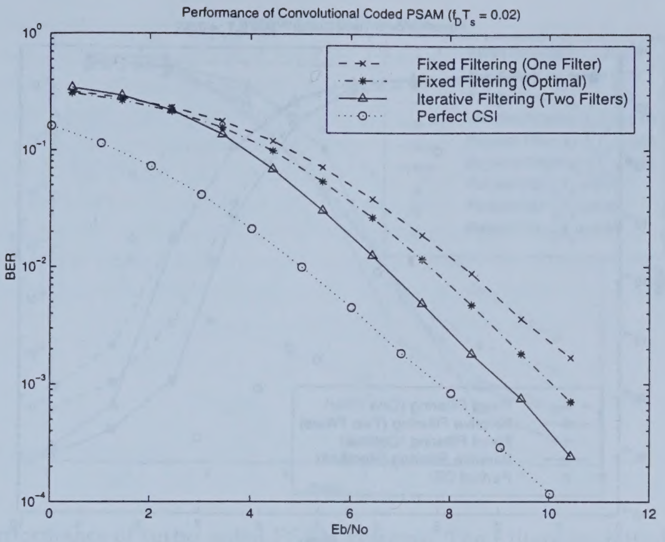


Figure 3: Convolutional coded PSAM performance comparison when the PSFs are matched to the channel condition. The optimal fixed-filter scheme has different PSFs for different symbol positions.

ratio of 10 data symbols per pilot symbol. Therefore, we had $M = 11$ symbols per PSAM frame; and each block of transmission contained 206 PSAM frames.

4.1 Convolutional Coded PSAM

In Fig. 2, we compare the performances of the proposed scheme and the conventional PSAM system with convolutional code. For both schemes, only one PSF is used. The PSF has 10 taps and was optimized for the symbol positioned at the middle of a PSAM frame for the channel condition $E_b/N_0 = 6$ dB and $f_D T_s = 0.02$, where T_s is the symbol duration. The proposed scheme uses an additional ASF which has also 10 taps covering the nearest 10 symbols adjacent to the symbol under consideration. The ASF was again optimized initially for $E_b/N_0 = 6$ dB and $f_D T_s = 0.02$. For both systems, the decoding takes 8 iterations. The conventional PSAM, however, does not have ASF and iterative filtering. We note that the step-size parameters μ_p and μ_a of the iterative filtering were determined empirically since there is no analytical formula describing the behavior from iteration to iteration.

From Fig. 2 we can see that the iterative-filter scheme improves the performance. For all fading conditions its gain over the conventional PSAM system is larger than 2 dB. Compared to the reference cases with perfect CSI, at 10^{-4} bit error rate (BER) the proposed scheme needs only about 1 dB more power, out of which 0.4 dB is consumed by the pilot symbols. It is also worth mentioning that for the conventional PSAM, once the channel fading rate goes beyond the rate for which the PSF was optimized, the performance becomes extremely unacceptable. This result is consistent with the results shown in [3]. While for the proposed scheme, the performance approaches that of the perfect CSI case within 1 dB even when the filters were initially mismatched.

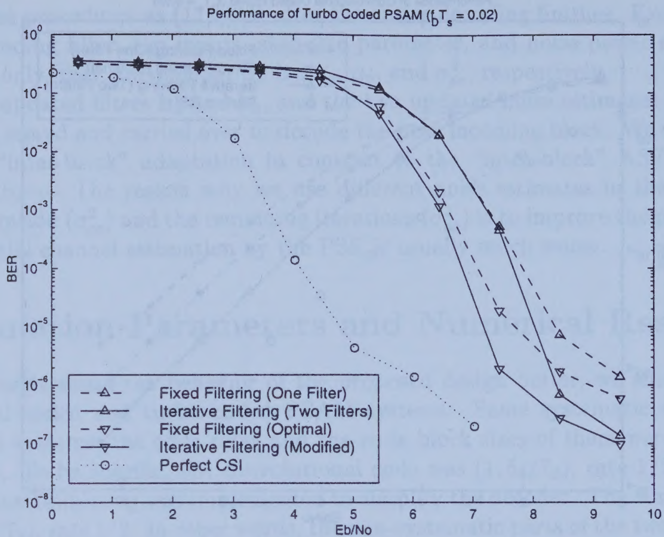


Figure 4: Turbo coded PSAM performance comparison when the PSFs are matched to the channel condition. The modified iterative-filter scheme has different PSFs and ASFs for different symbol positions. Each ASF uses 4 nearest pilot symbols and 6 nearest data symbols as inputs.

From the simulation results we conclude that the filter adaptation not only helps with the filter mismatch, it also improves the performance when there is no mismatch. This is shown clearly in Fig. 3 where we plot only the cases without filter mismatch. In this figure we can see that the iterative-filter scheme performs even better than the conventional PSAM system with 10 optimal PSFs applied to different symbol positions in a PSAM frame. Finally, we would like to point out that the proposed scheme has higher BER than the conventional PSAM system when E_b/N_0 is low. This phenomenon can be attributed to the adaptation induced error propagation; and it may concern the application of iterative filtering to a turbo coded system.

4.2 Turbo Coded PSAM

The single-PSF-single-ASF iterative scheme was applied to a turbo coded PSAM system. Due to the abrupt drop of the turbo code BER after the low E_b/N_0 region, the iterative-filter scheme shows hardly any improvement before the BER is low. As we can see from Fig. 4 in which we compare only the cases without filter mismatch, the curve of the iterative scheme with single PSF and single ASF departs from that of the conventional PSAM system only after 10^{-3} BER. Between the flat BER region and 10^{-3}, the performance is basically dominated by the turbo coding effect. For this reason we tried to shorten the error propagation region and expected that the advantage of the iterative filtering would happen earlier. We did so by using optimal filters for different symbol positions, and more importantly, including more pilot symbols in the ASFs. In other words, 10 PSFs, each optimized for a symbol position; and 10 ASFs, each operating on a symbol position and taking the nearest 6 data symbols and 4 pilot symbols as inputs, were used. The

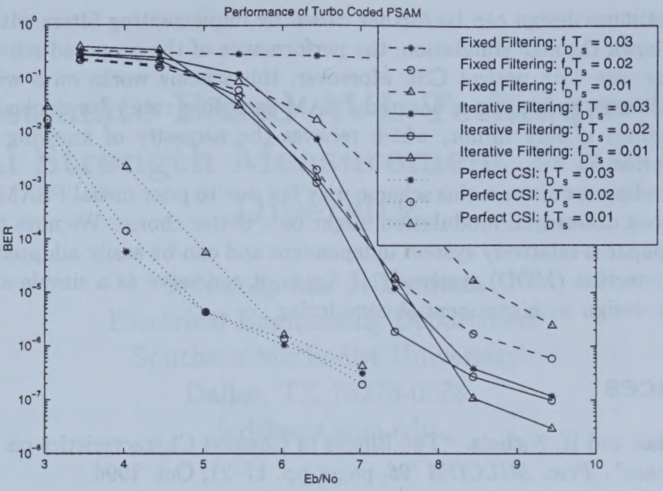


Figure 5: Performance of turbo coded PSAM systems. The Filters are initially optimized for $f_DT_s = 0.02$ and $E_b/N_0 = 6$ dB

simulation result of this modified iterative-filter scheme is shown in Fig. 4. It is clear that the error propagation region is shortened; and the curve of the iterative-filter scheme drops earlier than that of the conventional PSAM with optimal PSFs. At the BER of 10^{-6}, the proposed scheme outperforms the conventional PSAM system by 1 dB; and it lags behind the perfect CSI case by 1.5 dB. From Fig. 4 we also notice that the performances of the two iterative-filter schemes approach each other at high E_b/N_0. Depending on the BER requirement, one might need only the single-PSF-single-ASF scheme.

In Fig. 5 we compare the performances of the turbo coded schemes (optimal fixed-filter and modified iterative-filter) under different fading conditions. All the filters were again initially optimized for $E_b/N_0 = 6$ dB and $f_DT_s = 0.02$ for each symbol position. For all cases, the proposed scheme is at least 1 dB better than the conventional PSAM system. The BER floor of the proposed scheme is also lower than that of the conventional scheme. According to this figure and the results shown in Fig. 2, we expect that the performance of the proposed scheme will eventually be within 1 dB of that of the perfect CSI case. When the channel fading rate is higher than where the PSFs were optimized at, again, the performance of the conventional PSAM system is unacceptable. Using turbo codes does not change the situation at all.

5 Conclusion

In this paper, we proposed a joint iterative filtering and decoding scheme for coded PSAM systems. The proposed scheme lets iterative decoding and PSAM filtering benefit from each other. To be specific, the decoding part gets better channel gain estimation from the adaptive PSAM filtering; while the filtering part uses the decoding reliability information to improve its tap weights. In contrast to the optimum joint design of channel estimation and decoding, this scheme needs only linear computational complexity and least modifications from the standard turbo decoders and PSAM filters. Its performance lag to the optimum design can be further closed by implementing filters with more taps.

As was shown through simulation, the performance of the proposed scheme is within 1.5 dB of the case with perfect CSI. Moreover, this scheme works on a wider range of channel conditions by having a boosted PSAM sampling rate. Its dependence on the fading conditions is very minor, which relieves the necessity of knowing the channel condition a priori.

For fast fading conditions, this scheme may fail due to poor initial PSAM filtering. In these situations differential modulation might be a better choice. We note that the concept in this paper is relatively system independent and can be easily adopted by multiple differential detection (MDD) systems[11]. Again, it can serve as a simple alternative to the optimum design with exponential complexity.

References

[1] M. Jordan and R. Nichols. "The Effects of Channel Characteristics on Turbo Code Performace". *Proc. MILCOM '96*, pages pp. 17–21, Oct. 1996.

[2] M. C. Valenti and B. D. Woerner. "Performance of Turbo Codes in Interleaved Flat Fading Channels with Estimated Channel State Information". *Proc. VTC '98*, pages pp. 66–70, May 1998.

[3] J. K. Cavers. "An Analysis of Pilot Symbol Assisted Modulation for Rayleigh Fading Channels". *IEEE Transactions on Vehicular Technology*, Vol. 40(No. 4):pp. 686–693, Nov. 1991.

[4] C. Douillard et al. "Iterative Correction of Intersymbol Interference: Turbo-Equalization". *European Transactions on Telecommunications*, Vol. 6:pp. 507–511, Sep.-Oct. 1995.

[5] I. D. Marsland and P. T. Mathiopoulos. "Multiple Differential Detection of Parallel Concatenated Convolutional (Turbo) Codes in Correlated Fast Rayleigh Fading". *IEEE Journal on Selected Areas in Communications*, Vol. 16(No. 2):pp. 265–275, Feb. 1998.

[6] William C. Jakes. *"Microwave Mobile Communications"*. Wiley, New York, 1974.

[7] E. K. Hall and S. G. Wilson. "Design and Analysis of Turbo Codes on Rayleigh Fading Channels". *IEEE Journal on Selected Areas in Communications*, Vol. 16(No. 2):pp. 265–275, Feb. 1998.

[8] P. Robertson, E. Villebrun, and P. Hoeher. "A Comparison of Optimal and Sub-Optimal MAP Decoding Algorithms Operating in the Log Domain". *Proc. IEEE ICC '95*, pages pp. 1009–1013, Jun. 1995.

[9] J. Hagenauer and P. Hoeher. "A Viterbi Algorithm with Soft-Decision Outputs and Its Applications". *Proc. IEEE GLOBECOM '89*, pages pp. 1680–1686, Nov. 1989.

[10] Simon Haykin. *"Adaptive Filter Theory, 3rd Ed."*. Prentice Hall, New Jersey, 1996.

[11] D. Divsalar and M. K. Simon. "Multiple-Symbol Differential Detection of MPSK". *IEEE Transactions on Communications*, Vol. 38:pp. 300–308, Mar. 1990.

ML Estimate Based AVL Technique Using Signal Strength Measurement With Time Diversity

I. Nomay and J. Dunham
Electrical Engineering Department
Southern Methodist University
Dallas, TX 75275-0338
jgd@seas.smu.edu

Abstract

By the year 2001 FCC regulations require PCS and cellular vendors to provide location information of every mobile user within an accuracy of 125 meters 67% of the time [1] in a combination of varaity of operating channels. However such information is useful not only in Wireless 911 as stated by the FCC, but also for better call execution. For example, exact location of a Mobile Unit (MU) can be utilized to reduce the number of unnecessary hand-offs and paging cost.

In this paper we present an Automatic Vehicle Location (AVL) technique based on signal strength measurements. A Circular Trileteration (CT) strategy is analyzed and used through a Maximum Likelihood Estimation (MLE) process. The prediction of signal strength from a base station to a MU will be based on the fourth law of power which can be viewed as a simplification of Okumara-Hata model. In addition, time diversity with selection combining is proposed as a mean of combating deep fades in the channel.

Preliminary results where the channel is assumed to be multi-path with thermal noise at the receiver produced a location error of 150 meters in a constellation of three base stations apart from each other by three kilometers. However, for an AWGN channel the location error is around 30 meters.

1 Introduction

An AVL system is a collection of electronic devices used to acquire information about the location of a MU. As mentioned in [2], AVL techniques can be grouped in the following categories: Dead Reckoning [2], Proximity [2], Map Matching [3], Adaptive Antenna [4], and Radio-location [10].

A radio-location approach consists of two stages. First, direct measurements are made of a radio signal traveling between the MU and a number of fixed location stations. These measurements are then used on a known geometric configuration to solve a set of nonlinear equations to determine the unknown location. Three types of measurements are usually used, Time of Arrival, Angle of Arrival (ToA, AoA) [6] and Signal Strength (SS). GPS is a well known system that uses ToA measurements [7].

A linear approximation of the set of location equations using a Taylor series expansion is presented in [8]. A location mean error of 663 feet and 169 meters are the results of two independent researches [9, 10] respectively. These two results are based on a Least Square Error Estimate (LSEE) process using the measurements of four sensors in a circle of radii about two miles in a multi-path channel. A MLE process was presented in [11] in order to track an object in noisy environment with the variance of the MLE error presented in [12]. SS measurements was used in [13] with an accuracy of 1000 foot for a spacing distance of 1.5 miles between any two base stations in a multi-path channel. The shadowing effect on the transmitted signal has been treated in [14]. Neural Networks and Fuzzy Logic methods are proposed in [15] with accuracy of 200 meters.

2 SS Measurements Model

Assume that a mobile unit is located within a constellation of three base stations with each base station transmitting a pilot signal $P_t(t)$ in watts. According to the fourth law of power [17],

$$P_{ri}(t) = \frac{P_{ti}(t)\ G_t\ G_r\ (h_t\ h_r)^2}{r_i^4(t)} \tag{1}$$

where $P_{ri}(t)$ and $P_{ti}(t)$ are the received power (at the mobile unit end) and transmitted power from base station i, respectively; G_r and G_t are the receiver and the transmitter antennas' gains, respectively; h_t and h_r are the transmitter and receiver antennas' height, respectively; and $r_i(t)$ is the distance between base station i and the mobile unit. Eq. (1) could be viewed as a simplified Okumara-Hata model.

3 Signal Model and Circular Trileteration (CT) Statistical Approach

The CT approach is based on the fact that there is a unique intersection point for three non co-linear circles. Such a point represents the exact MU's location. The three circles' radii are the estimated distances between each base station's center and the MU's location. The prediction model shown in Eq. (1) is the basis for the distance estimations.

For an outdoor cellular system with a multi-path and thermal noise at the receiver the received signal would be in the form

$$v_m(t) = a(t)v_r(t) + n(t) \tag{2}$$

where $v_m(t)$ is the measured voltage (observation), $v_r(t)$ is the received voltage, $a(t)$ is a Rayleigh distributed random variable with a variance of σ_a^2, and $n(t)$ is a zero mean Gaussian process with a variance of σ_n^2. The received voltage can be related to the distance $r(t)$ traveled by the signal from base station to the mobile unit using Eq. (1), *i.e.*,

$$v_r(t) = \frac{k}{r^2(t)} \tag{3}$$

where k is a constant depends on the system parameters, such as antennas' highs and gains and frequency used. Substituting Eq. (3) into Eq. (2), gives

$$r(t) = \sqrt{\frac{a(t)\ k}{y(t)}} \tag{4}$$

where $y(t) = |v_m(t) - n(t)|$, $y(t) > 0$ with a Probability Density Function (PDF) of [19]

$$f(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}(\exp\,(- \frac{(y - v_m)^2}{2\,\sigma_y^2}) \;+ \exp\,(- \frac{(- y - v_m)^2}{2\,\sigma_y^2}))\quad 0 < y < \infty. \tag{5}$$

Letting $x = a/y$ yields

$$f_R(r) = f_X(\frac{r^2}{k})\;\frac{2\,r}{k} \tag{6}$$

where [20]

$$f(x) = \int_{-\infty}^{\infty}\; |v|\; f_A(vx)\; f_Y(v)\; dv. \tag{7}$$

Solving for $f_X(x)$ and transforming random variables gives $f_R(r)$ which is found to be [19]

$$f(r) = \frac{r^3}{2k^4\sigma_a^4\sqrt{2\sigma_y^10}}\sqrt{(\frac{2k^2\sigma_a^2\sigma_y^2}{r^4\sigma_y^2 + k^2\sigma_a^2})^5}\;.(k^2\sigma_a^2(\sigma_y^2{+}V_m^2){+}r^4\sigma_y^4)\;.\exp\,(\frac{V_m^2}{2\sigma_y^2}(\frac{k^2\sigma_a^2}{r^4\sigma_y^2 + k^2\sigma_a^2} - 1)) \tag{8}$$

and the cumulative density function (CDF) is [19],

$$F(r) = 1 - \sqrt{\frac{k^2\sigma_a^2}{r^4\sigma_y^2 + k^2\sigma_a^2}}\;.\exp\,(\frac{V_m^2}{2\sigma_y^2}(\frac{k^2\sigma_a^2}{r^4\sigma_y^2 + k^2\sigma_a^2} - 1)) \tag{9}$$

3.1 One Base Station

Let $f(r_i, \theta)$ to be the joint PDF of R_i and θ, where (r_i, θ) are the coordinates of the MU location. Since R_i and θ are mutually independent RV's,

$$f(r_i, \theta) = f(r_i)\; f(\theta) \tag{10}$$

where $f(r_i)$ is given by Eq. (8) and $f(\theta)$ is the probability density function of the angle θ between the line connecting base stations i and j and the MU location.

In the case when only one base station measurement is available, $f(\theta)$ is a uniform density function on $[0, 2\pi]$, *i.e.*,

$$f(\theta) = \begin{cases} \frac{1}{2\pi} & 0 < \theta \le 2\pi \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Fig. 1 shows $f_{R_i,\theta}(r_i, \theta)$ in the case where all locations have the same probability.

3.2 Two Base Stations

If two base stations are involving in the location determination process, (R_i, R_j), the values of θ will shrink to only two possibilities whose values are given by

$$\theta = \begin{cases} |\cos^{-1}(\frac{r_i^2 + d_{ij}^2 - r_j^2}{2\;r_i\;d_{ij}})| \\ -\;|\cos^{-1}(\frac{r_i^2 + d_{ij}^2 - r_j^2}{2\;r_i\;d_{ij}})| \end{cases} \tag{12}$$

where d_{ij} is the distance between base stations i and j.

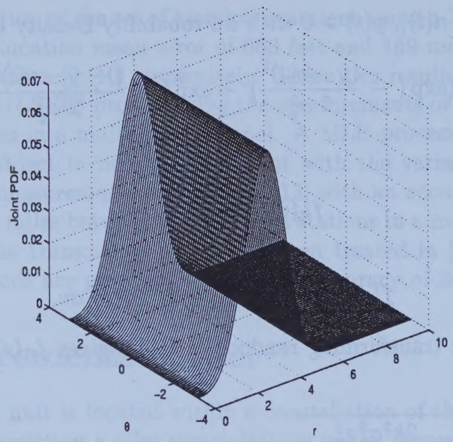


Figure 1: Joint PDF of r_i and θ (case of one base station).

In this case the joint PDF $f(r_i, \theta)$ can be evaluated using the chain rule, *i.e.*,

$$f(r_i, r_j, \theta) = f(r_i)\ f(r_j|r_i).\ f(\theta|r_i, r_j) \tag{13}$$

where $f(r_j|r_i)$ is the conditional PDF of R_j given R_i and is found to be

$$f(r_j|r_i) = \frac{f(r_j, r_i)}{f(r_i)} = fn(r_j) \tag{14}$$

where $fn(r_j)$ is the normalized version of $f(r_j)$ over the interval $[|d_{ij} - r_i|, d_{ij} + r_i]$,

$$fn(r_j) = \frac{f(r_j)}{\int_{|d_{ij}-r_i|}^{d_{ij}+r_i} f(r_j)\ dr_j}\ , \tag{15}$$

and $f(\theta|r_i, r_j)$ is a discrete density function that has only two values occurring at $\pm\theta$,

$$f(\theta|r_i, r_j) = \delta_\theta(r_i, r_j) = \begin{cases} \frac{1}{2} & |\cos^{-1} \frac{r_i^2 + d_{ij}^2 - r_j^2}{2\ r_i\ d_{ij}}| \\ \frac{1}{2} & -|\cos^{-1} \frac{r_i^2 + d_{ij}^2 - r_j^2}{2\ r_i\ d_{ij}}| \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Therefore, $f(r_i, \theta)$ will be

$$\begin{aligned} f(r_i, \theta) \ &= \int_{R_j} f(r_i, r_j, \theta)\ dr_j \ = f(r_i) \int_{|d_{ij}-r_i|}^{d_{ij}+r_i} fn(r_j)\ f(\theta|r_i, r_j)\ dr_j \\ &= f(r_i)\ z_1^{'}\ (\frac{fn_{R_j}(z_1)}{2z_1} - \frac{fn_{R_j}(z_2)}{2z_2}). \end{aligned} \tag{17}$$

where

$$z_k = \sqrt{r_i^2 + d_{ij}^2 \mp 2r_i d_{ij}\ \cos\ \theta}, \quad k = 1, 2 \tag{18}$$

Now $f(r_i, \theta)$ is completely defined and is shown in Fig. 2. It is noticeable that there are two peaks in this case and these two peaks represent the two possible locations of the mobile terminal. These two locations are symmetric around a reference line that connects base stations i and j. Later, the reference line d_{ij} will be adjusted to the x-axis.

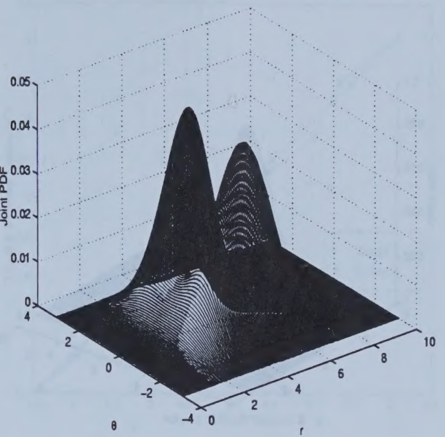


Figure 2: Joint PDF of r_i and θ (case of two base stations).

3.3 Three Base Stations

Third base stations measurement will virtually eliminate one of the previous two possible solutions, leaving the one which represents the exact location. In this case the joint PDF of all observations (R_i, R_j, R_k) will be

$$f(r_i, r_j, r_k, \theta) = f(r_i)\ f(r_j|r_i)\ .f(\theta|r_i, r_j)\ .f(r_k|r_i, r_j, \theta) \tag{19}$$

where $f(r_i)$, $f(r_j|r_i)$, $f(\theta|r_i, r_j)$ are given by Eq.'s (8), (14), and (16), respectively, and $f(r_k|r_i, r_j, \theta)$ is found to be

$$f(r_k|r_i, r_j, \theta) = \begin{cases} p & r_k^1 \\ (1-p) & r_k^2 \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where

$$r_k^i = \sqrt{(y_k - y_i)^2\ +\ (x_k - x_i)^2}, \quad i = 1, 2 \tag{21}$$

and

$$p = \frac{f(r_k^1)}{f(r_k^1) + f(r_k^2)}. \tag{22}$$

Therefore $f(r_i, \theta)$ will be

$$\begin{aligned} f(r_i, \theta) &= \int_{R_j} \int_{R_k} f(r_i, r_j, r_k, \theta) dr_j\ dr_k \\ &= f(r_i) \int_{|d_{ij} - r_i|}^{d_{ij} + r_i} \int_0^\infty f(r_j|r_i)\ f(\theta|r_i, r_j).\ f(r_k|r_i, r_j, \theta)\ dr_j\ dr_k \\ &= f(r_i) \int_{|d_{ij} - r_i|}^{d_{ij} + r_i} f(r_j|r_i)\ f(\theta|r_i, r_j).\ \int_0^\infty f(r_k|r_i, r_j, \theta)\ dr_j\ dr_k \\ &= f(r_i)\ z_1'\ (\frac{fn_{R_j}(z_1)\ p}{2z_1} - \frac{fn_{R_j}(z_2)(1-p)}{2z_2}) \end{aligned} \tag{23}$$

This density function is shown in Fig. 3. It is clear now that one of the peaks in Fig. 2 has been strongly suppressed and the location of the mobile terminal has been localized to a single area.

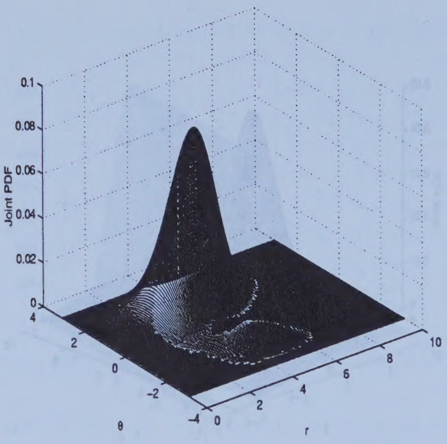


Figure 3: Joint PDF of r_i and θ (case of three base stations).

Table 1: Simulation parameters

Parameter	Value
k	60
Rayleigh Variance (σ_a^2)	1.0
AWGN Variance (σ_n^2)	1.0
Base Station i coordinates	(2,2)
Base Station j coordinates	(7,5)
Base Station k coordinates	(3,6)
No. of Observations / Base Station	10

4 ML Estimate and Simulation Results

A simulation of the location estimation process was performed using the parameters shown in Table 1. In this simulation the MU is assumed to move in a constant velocity along a straight line inside a constellation of three cells separated from each other by three kilometers. A ML estimate is performed at a specific segment of time by computing the values of r_i and θ that have maximum probabilities.

ML estimate of the distance r_i is found by solving the log-likelihood equation

$$\frac{\partial}{\partial r_i}\ln(f_{R_i}(r_i)) = 0 \tag{24}$$

and so $\hat{r}_{MLE}$ is found to be [19]

$$\hat{r}_{MLE} = |real(\frac{(9c_2)^{\frac{1}{4}}}{3\sigma_n^2})| \tag{25}$$

where

$$c_2 = \frac{k^2\sigma_a^2\sigma_n^2(4\sigma_n^4 + 7v_m^4 + c_1^{\frac{2}{3}} - 3v_m^2 c_1^{\frac{1}{3}} - c_1^{\frac{1}{3}}\sigma_n^2)}{c_1^{\frac{1}{3}}} \tag{26}$$

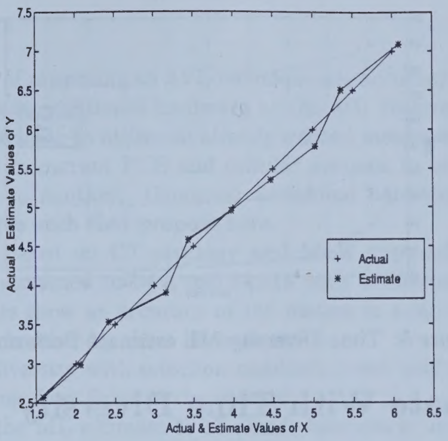


Figure 4: Actual and estimate paths (Km.).

and

$$c_1 = 8\sigma_n^6 + 3\sigma_n^2 v_m^4 - 18v_m^6 + v_m^2(-288\sigma_n^6(\sigma_n^2 + v_m^2) - \sigma_n^2 v_m^4(579\sigma_n^2 + 108v_m^2) - 19v_m^8)^{\frac{1}{2}}. \quad (27)$$

Similarly, the ML Estimate of θ is found by solving,

$$\frac{\partial}{\partial\theta}\ \ln(f(\theta)) = 0. \quad (28)$$

Using Eq. (23), $f_\theta(\theta)$ is found to be

$$f(\theta) = \frac{f n_{R_j}(z_1)\ z_1^{'}}{2\ z_1}. \quad (29)$$

Because r_i and θ are independent, the two values of θ have the same magnitude. Substituting Eq.'s (8), and (15) into Eq. (29) gives

$$\begin{aligned} f(\theta) \ &= \ \frac{k_1 z_1^2 z_1^{'}}{2k^4\sigma_a^4\sqrt{2\sigma_y^10}} \sqrt{(\frac{2k^2\sigma_a^2\sigma_y^2}{z_1^4\sigma_y^2 + k^2\sigma_a^2})^5} \ \times \ (k^2\sigma_a^2(\sigma_y^2 + V_m^2) + z_1^4\sigma_y^4) \\ & \times \ \exp(\frac{V_m^2}{2\sigma_y^2}(\frac{k^2\sigma_a^2}{z_1^4\sigma_y^2 + k^2\sigma_a^2} - 1)) \ \ 0 < \theta < \pi \end{aligned} \quad (30)$$

where $k_1 = 1/(2\int_{|d_{ij}-r_i|}^{d_{ij}+r_i} \ f_{R_j}(r_j)\ dr_j)$. The solution of Eq. (28) is equivalent to finding the roots of a polynomial of degree >4. This has no analytical solution and has to be evaluated numerically.

The decision of which position to chose is based on its probability, *i.e.*,

$$\theta = \begin{cases} +\ |\cos^{-1}\frac{r_i^2+d_{ij}^2-r_j^2}{2\ r_i\ d_{ij}}| & p > (1-p) \\ -\ |\cos^{-1}\frac{r_i^2+d_{ij}^2-r_j^2}{2\ r_i\ d_{ij}}| & \text{otherwise.} \end{cases} \quad (31)$$

Fig. 4 shows both the actual and the estimate paths of the mobile terminal obtained by MLE estimation process.

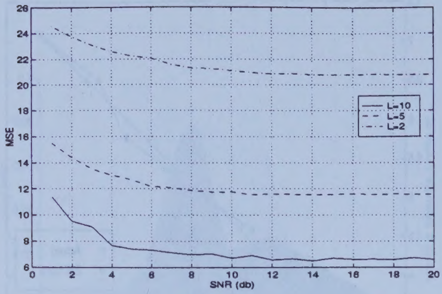


Figure 5: Time Diversity ML estimate Performance

5 ML Estimate With Time Diversity

Diversity is an efficient tool for combating deep fades on channels because multiple independent paths are much less likely to experience deep fades simultaneously than a single channel. Achieving independence could be obtained through receiving replicas of the original signal. These replicas could be in the form of time, frequency, spatial, or polarization diversity. In case of micro-diversity, where the diversity is needed at the MU end, a time separation between multiple replicas of the signal is easily achieved. In order to achieve l independently fading versions of the signal it is necessarily to transmit the signal in different time slots. The separation between successive time slots equals or exceeds the coherence time of the channel [16]. For example, the separation between successive time slots must exceed 8.3 ms for a mobile at 72 km/h, operating at 900 MHz,[17].

The actual distance traveled is the one represented by the Line Of Sigh (LOS) path which is assumed not to be visible in a Raleigh modeled channel. Therefore, the receiver is forced to select the branch that represents the shortest distance instead. let $R_d = min(r_1, r_2, ..., r_l)$ where l is the number of independent diversity branches. Assume that $r_1, r_2, ..., r_l$ are i.i.d. r.v. Therefore, the CDF $F(r_D)$ will be,

$$F(r_d) = p[R_d \leq r_d] = 1 \ - \ [(1 - F_{R1}(r1))(1 - F_{R2}(r2)) \, ..(1 - F_{Rl}(rl))]. \tag{32}$$

Let $\sigma_{r_1}^2 = \sigma_{r_2}^2 = .. = \sigma_r^2$, then,

$$F(r_d) = 1 \ - \ [(1 \ - \ F_R(r))^l] \tag{33}$$

and so the PDF of r_d will be,

$$f(r_d) = l \ f_R(r) \ (1 \ - \ F_R(r))^{(l-1)}. \tag{34}$$

Substituting Eqs. (8), (9) into Eq. 34 and using the equality given in [21] yield

$$f(r_d) = \frac{2 \ l \ r_d^3 \ k^l \ \sigma_a^{\frac{l}{2}}}{(r_d^4 \sigma_y^2 + k^2 \sigma_a^2)^{\frac{l+4}{2}}} \, .(k^2 \sigma_a^2 (\sigma_y^2 + V_{md}^2) + r_d^4 \sigma_y^4) \ . \left\{ \exp \left(\frac{V_{md}^2}{2\sigma_y^2} (\frac{k^2 \sigma_a^2}{r_d^4 \sigma_y^2 + k^2 \sigma_a^2} - 1) \right) \right\}^l \tag{35}$$

Finally, solving for the likelihood equation gives the ML estimate of r_d which has no analytical solution and has to be evaluated numerically.

Figure (5) shows the effect of diversity in minimizing the Mean Square Error (MSE).

6 Conclusion

The two main criterion of proposing an AVL technique are accuracy and cost effectiveness. A GPS based AVL needs additional hardware at the MU receiver end plus the cost of licensing. It is a wise decision to utilize an already existed measurements. Signal strength measurement are used in current PCS and cellular systems in order to hand-off a MU from one base station to another. Hence no additional hardware will be added for a signal strength technique such that propose here.

A statistical AVL based on CT strategy and MLE approach is presented in this paper. The channel is assumed to be a multi-path with a thermal noise at the receiver end. Preliminary results show an accuracy of 150 meters in a constellation of three base stations apart from each other by three kilometers.

In addition, time diversity with selection combining was analyzed and showed to be an efficient in combating deep fades in the channel.

A modification to the ML estimate to eliminate biasness is under research as well as the development of more robust tracking algorithms.

References

[1] M. Aatique, G. Mizusawa and B. Woerner, "Performance of Hyperbolic Position Location Techniques for Code Division Multiple Access," *The 9th Int. Conf. on Wireless Comm.*, Vol. 1, pp. 261-274, July 1997.

[2] S. Rite and J. McCoy, "Automatic Vehicle Location - An Overview," *IEEE Trans. on Veh. Tech.*, Vol. VT-26, No. 1, pp. 7-11, February 1977.

[3] F. Morisue and K. Ikeda, "Evaluation of Map-Matching Techniques," *Proc. IEEE Veh. Nav. and Inf. Sys. (VNIS)*, Toronto, Canada, pp. 23-8, 1989.

[4] R. Davis, L. Brennan and L. Reed, "Angle Estimation with Adaptive Arrays in External Noise Fields," *IEEE Trans on Aero. and Elect. Systems*, Vol. AES-12, No. 2, pp. 179-86, March 1976.

[5] H. Hashemi, "Pulse Ranging Radio location Technique and Its Application to Channel Assignment in Digital Cellular Radio," *Proc. IEEE 41st Veh. Tech. Conf. (VIC)*, St. Louis, MO, pp. 675-80, 1991.

[6] J. Caffery Jr. and G. Stuber, "Subscriber Location in CDMA Cellular Networks", *IEEE Trans. on Veh. Tech.*, Vol. VT-47, No. 2, pp. 406-416, May 1998.

[7] Elliott Kaplan, *Understanding GPS Principles and Applications*, Artech house, Inc., 1996.

[8] W. Foy, "Position Location Solutions By Taylor-Series Estimation," *IEEE Trans. on Aero. Elect. Systems.* Vol. AES-12, no. 2, pp. 187-93, March 1976.

[9] G. Turin, W. Jewell and T. Johnston, "Simulation of Urban Vehicle-Monitoring Systems," *IEEE Trans. on Veh. Tech.*, Vol. VT-21, pp. 1-9, February 1972.

[10] H. Hashemi, "Pulse Ranging Radio location Technique and Its Application to Channel Assignment in Digital Cellular Radio," *Proc. IEEE 41st Veh. Tech. Conf. (VIC)*, St. Louis, MO, pp. 675-80, 1991.

[11] L. Favalli, P. Gamba, T. Gatti and A. Mecocci, "Multi-radar Data Fusion for Object Tracking and Shape Estimation," *Signal Processing*, Vol. 48, pp. 235-9, 1996.

[12] A. Schatzberg, A. Devaney and A. Witten, "Estimating Target Location from Scattered Field Data," *Signal Processing*, Vol. 40, pp. 227-37, 1994.

[13] W. Figel, N. Shepherd and W. Trammell, "Vehicle Location by a Signal Attenuation Method," *IEEE Trans on Veh. Tech.*, Vol. VT-18, no. 3, pp. 105-9, November 1969.

[14] M. Hata and T. Nagatsu, "Mobile Location Using Signal Strength Measurements in a Cellular System," *IEEE Trans. on Veh. Tech.*, Vol. VT-29, No. 2, pp. 245-51, May 1980.

[15] H. Song, "Automatic Vehicle Location in Cellular Communications Systems," *IEEE Trans. on Veh. Tech.*, Vol. VT-43, No. 4, pp. 902-8, November 1994.

[16] J. Proakis, *Digital Communications*, 3rd Edition, McGraw Hill, 1995.

[17] M. Yacoub, *Foundations of Mobile Radio Engineering*, CRC Press,Inc, 1993.

[18] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd Edition, McGraw Hill, 1991.

[19] I. Nomay, "Mobility Tracking Using AVL Technique," Ph.D. Dissertation, Electrical Engineering Dept., Southern Methodist University, In progress.

[20] A. Dorogovtsev, D. Silvestrov, A. Skorokhod and M. Yadrenko, "Mathmatical Monographs," Vol. 163, AMS Providence, RI, 1997.

[21] I. Gradshteyn and I. Ryzhik, *Tables of Integrals, Series, and Products*, Academic Press, 1980.

Radio network optimization by means of neural networks *

Arnaud Berny
IRIN, Université de Nantes
2, rue de la Houssinière
BP 92208
44322 Nantes cedex 3 France
ECTIA
1, rue de la Noë
BP 92119
44321 Nantes France
Arnaud.Berny@ectia.ec-nantes.fr

Abstract

We propose new algorithms for designing and optimizing cellular radio networks. The aim is to cover a given area with cells whereas minimizing the cost of the network. Interference and traffic issues are also taken into account. One of the core ideas is that potential sites are considered as interacting particles. We show how sites are implicitly interacting through cells. The relation between antenna parameters and cells is learnt by radial basis function (RBF) neural networks. This relation is used to estimate antenna parameters whose cell best matches a cover hole around a site. The reduction of dimensionality of cell images is achieved with a convolutional neural network introduced in this paper.

1 Introduction

There has been an increasing demand on cellular radio networks for the last few years and this trend does not seem to slow down. This demand led operators to seek automated tools for the design and optimization of networks, in particular for frequency assignment. Now they are looking for automated cell planning tools which are able to manage many objectives and constraints at the same time. This radio network optimization problem shows both combinatorial and continuous sides, which makes it difficult.

The continuous part, i.e., the choice of antenna parameters, is addressed with radial basis function neural networks. It is viewed as an image matching problem which is to find a cell close to a given cover hole. Images are preprocessed by a convolutional neural network in order to reduce the complexity of subsequent stages.

The combinatorial part of the optimization is addressed with statistical physics tools used in spin-glass models or stochastic image restoration. Sites are seen as interacting particles and a collective behavior is sought in order to reach a good radio network.

---

*This work is funded by the EC under contract No. 23243 (ARNO project).

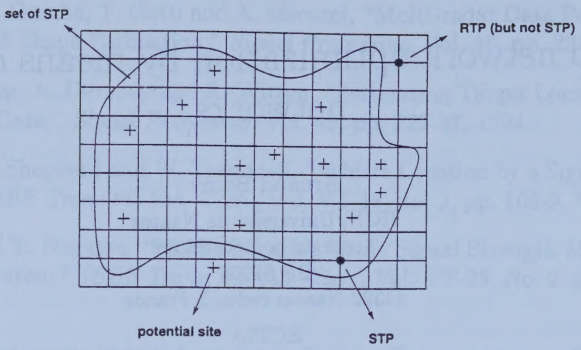


Figure 1: Basic sets of points in the area to cover.

The paper is organized as follows. Section 2 introduces the radio network optimization problem. Section 3 details the antenna parameter estimation, including the convolutional and RBF neural networks. Finally, Section 4 presents the stochastic algorithm which updates site configurations.

2 The radio network optimization problem

2.1 Basic sets

The area to be covered is sampled on a rectangular grid whose vertices are called reception test points (RTP). Only a subset of RTP needs to be covered; those points are called service test points (STP). Those sets are represented on Fig. 1.

The area contains a set of (fixed) potential sites, each of which can hold up to three base stations — a base station is an antenna plus one or more transmitters. There are two types of antenna: directional ones and omnidirectional ones. Those types are exclusive on the same site, i.e., a site can hold one omnidirectional antenna or one to three directional ones. There exists different kinds of directional antenna but we restrict ourselves to only one.

The state of a base station is completely described by three parameters: power of the transmitter, azimuth and tilt. The last two ones give the orientation of the antenna. The state of an omnidirectional one is completely described by only one parameter, the power of the transmitter.

2.2 Propagation model

The propagation model is restricted in the sense that the only information available is the loss from any site to any RTP and the angle of incidence under which any antenna sees any RTP. This information precisely synthesizes many electromagnetic phenomena such as reflexion, absorption or diffraction.

In this model, only amplitude of the electromagnetic field is considered, more precisely its logarithm (expressed in dB). Most computations on signals thus reduce to additions (gain) and substractions (loss).

Each STP receives a set of field strenghs coming from the stations in the radio network. Only fields above a sensibility threshold t_s are considered. Those fields are ordered in descending order $f_1 > f_2 > \cdots > f_n > t_s$. A quality threshold t_q is

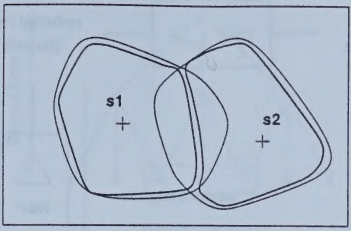


Figure 2: Difference between theoretical (thin line) and actual cells (thick line) for two stations on two neighbouring sites.

defined for each STP. It depends on the desired service (indoor, outdoor, car) and on the expected power of mobiles. If there exists at least one f_i above t_q the STP is covered. If $f_1 > t_q$ then the STP is assigned to the station that generated f_1 which is called "best server".

Each station may define a cell (provided its power is sufficient). There are indeed two different cells:

- a theoretical one, which is the set of STP where the signal is above the quality threshold t_q; it does not depend on any other station;
- an actual one, which is the set of STP where the signal is above t_q and also above any other signal; the station is the best server of this set. The actual cell depends on other stations.

The Fig. 2 illustrates this difference.

In order to ensure continuity of mobile communications across adjacent cells, radio networks provide a "handover" mechanism between adjacent cells so that an outgoing communication can be managed by at least one other station. In each cell there must exist "handover" STP belonging to at least two different (theoretical) cells, where the gap between field strenghs is lower than g_h. At most four fields may be candidate handover fields and the remaining ones are considered as interferer fields. In our notations handover fields are $f_2 \ldots f_5$ and interferer fields are $f_6 \ldots f_n$. We measure interference on an STP by the quantity $\sum_{i=6}^{n}(f_i - t_s)$.

2.3 Objectives, constraints and cost function

Traffic issues are central in the design of radio networks. We must ensure that traffic demand is served, both at a cell level and at a network level. Let us describe the way how traffic information is dealt with. A positive or null traffic quantum (expressed in Erlang) is assigned to each STP. The traffic covered by a cell is defined as the sum of traffic quanta over its STP. In order to serve this traffic demand, a station can hold one to seven transmitters — seven transmitters can serve a maximum of 43 Erlangs.

Another issue is cell topology. In order to limit handover between cells (management of the radio network) cells should consist in one connected component. This strong constraint is relaxed and "small" connected components are discarded.

The distinction between objectives and constraints is not definitive and depends on the priorities of the operator. Let us summerize the choices we have made in this model. For the constraints we have:

- cover: each STP must have a best server;

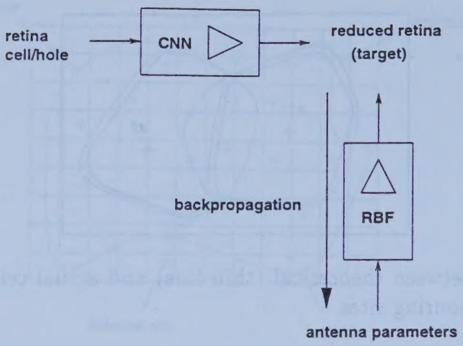


Figure 3: Architecture for antenna parameter estimation. CNN stands for convolutional neural network.

- the covered traffic of each cell must be under 43 Erlangs, i.e., the traffic capacity allowed by seven transmitters.
- each cell should consist in one connected component;

And for the objectives we have:

- minimize the number of used sites;
- minimize interference.

Only cover and traffic constraints are *explicitly* dealt with in our approach. Interference objective is taken into account during the parameter estimation stage. Nevertheless all constraints and objectives are combined in a linear cost function involved in the stochastic dynamics of site configurations.

3 Antenna parameter estimation

3.1 Introduction

Given a station and a cover hole, the aim of this stage is to estimate antenna parameters which lead to a theoretical cell close to the cover hole. This specification limits interference in the sense that cell overlap will be controlled.

The estimation is performed with two different kinds of neural networks. The first one reduces the complexity of images to be analyzed. It is a multilayer feedforward linear neural network that extracts most of information in original images and reduces their dimension. Its output is still two-dimensional. It can be thought of as an alternative to principal component analysis (PCA) neural networks. It is called in this paper a convolutional neural network.

The second kind of neural network is a radial basis function (RBF) neural network — its output is a linear combination of Gaussians. Its inputs are antenna parameters and its outputs are the outputs of the convolutional neural network. The estimation is performed by backpropagation to the inputs.

Image analysis is restricted to a square centered on a site, called *retina*. The retina dimensions are 48 × 48 square RTP. The retina is the input to the convolutional neural network. We insist upon the fact that images are binary images. The overall architecture is charted on Fig. 3.

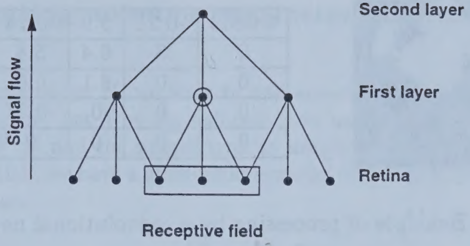


Figure 4: One-dimensional architecture of the convolution neural network.

3.2 Convolutional neural network

This neural network owes much to the work of Linsker [1]. Fig. 4 charts its architecture (one-dimensional for clarity). The neurons are organized in a pyramidal way. Each neuron takes its inputs from neurons situated in a small region in the previous layer (or the retina). This region is called a *receptive field.* One may link this architecture to multiresolution image analysis. An important parameter is the radius of the receptive field. Receptive fields may overlap.

Following [1], the main guideline in the design of the learning rule is that each neuron transmits most of information embodied in its receptive field. Let $\mathbf{x}$ be the receptive field vector of a neuron, y its scalar output and $\mathbf{w}$ its weight vector. We have $y = (\mathbf{x}|\mathbf{w})$ where $(\cdot|\cdot)$ is the scalar product. We write the learning rule as a differential equation:

$$\frac{d\mathbf{w}}{dt} = y(\mathbf{x} - y\mathbf{w}) \qquad (1)$$

This rule is implemented as a discrete time stochastic algorithm. The weight vector $\mathbf{w}$ converges to the first principal component of the underlying probability distribution $p(\mathbf{x})$ and maximizes the variance $\langle y^2 \rangle$ of the output signal as a measure of the transmitted information.

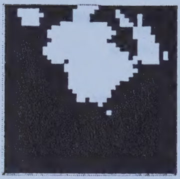
Let us describe the learning protocol. Layers are trained successively, starting from the input one to the output one. At each iteration a retina is sampled from its distribution — theoretical cell images generated from the uniform distribution of antenna parameters. This retina is propagated to the trained layer and only weights in this layer are updated using Eq. 1.

Fig. 5 charts a retina and its processing by a convolutional neural network. One advantage of this device over PCA is its low complexity. Let us consider a convolutional neural network with two layers of dimensions 16×16 and 4×4 taking its input from a 48×48 retina. Set both radii to 2. The number of scalar weights is 6400 to be compared with the 36864 weights of a linear transform (such as PCA) from retinas to outputs.

A deeper analysis of convolutional neural networks in comparison with PCA needs to be carried out. Nevertheless, this architecture proved to be a good preprocessing stage for the estimation of antenna parameters.

3.3 Learning cells

The convolutional neural network provides reduced retinas and enables their learning by neural networks within a reasonable complexity. Returning to our goal of estimating antenna parameters, in order to fill any cover hole, we might think of



0.0032	-0.21	3.9	-0.14	-0.0031
0	3	6.4	5.4	2.1
0	0	8.1	6.6	-0.14
0	0	0	-0.27	0
0	0	0	0	0

Figure 5: Example of processing by a convolutional neural network.

learning the relation: (reduced) retinas ↦ parameters. The drawback is that the class of cover hole images is much wider than the class of cell images and harder to learn. If a neural network trained with cell images is presented with a cover hole (whose distribution differs from the cell distribution) its outputs might have no meaning at all in terms of antenna parameters. Moreover the learning of the direct relation is likely to be *ill-posed*, since a given cell image may map to many parameter vectors (which generate the cell).

The reverse relation: parameters ↦ (reduced) retinas is easier to learn (it is a function). Backpropagation to the inputs provides a way to change parameters in order to make the corresponding cell closer to the target cover hole.

The device that learns reduced cells is a radial basis function (RBF) neural network (see [2]). The chosen RBF is the scaled Gaussian kernel $\exp(-t^2/\sigma^2)$. Basically, a RBF neural network implements a linear combination of Gaussian kernels. Let F be the function to learn:

$$\begin{array}{rccc} F: & \mathbb{R}^m & \longrightarrow & \mathbb{R}^p \\ & x & \longmapsto & F(x) \end{array}$$

Let (z_j) be a familly of n reference vectors in $\mathbb{R}^m$. Each z_j is assigned a RBF r_j at scale σ_j. Let us introduce some linear algebra notations:

$$\begin{aligned} r_j(t) &= \exp(-t^2/2\sigma_j^2) \\ R &= (\ldots r_j(||x - z_j||) \ldots)^t \\ F &= (\ldots f_i(x) \ldots)^t \\ A &= [a_{ij}] \end{aligned}$$

where R is an n-vector and A a $p \times n$ matrix. $()^t$ denotes matrix transpose. The output of the neural network is $\tilde{F} = AR$.

The learning rule derives from the minimization of the mean square error and is an "on-line" stochastic algorithm. It expresses as:

$$\begin{aligned} A(t+1) &= A(t) + \eta(F - AR)R^t \\ z_j(t+1) &= z_j(t) + \eta\frac{(x - z_j)}{\sigma_j^2} r_j(||x - z_j||) \sum_i a_{ij}\Delta_i \\ \sigma_j^2(t+1) &= \sigma_j^2(t) + \eta\frac{||x - z_j||^2}{\sigma_j^4} r_j(||x - z_j||) \sum_i a_{ij}\Delta_i \end{aligned}$$

where $\Delta_i = f_i(x) - \sum_j a_{ij} r_j(||x - z_j||)$ is the ith output error, $\sum_i a_{ij}\Delta_i$ is the backpropagated error to the jth RBF and η is the learning rate. Those expressions closely follow those of backpropagation for multilayer perceptrons.

3.4 Estimating antenna parameters by means of inversion

Given a cover hole retina H, we want to find antenna parameters x that generate the closest cell to H. Let h be the reduced cover hole retina of H. The estimated reduced cell is $\tilde{F}(x)$ and the square error is simply $E = ||h - \tilde{F}(x)||^2$. After the training of the RBF, we have a differentiable model of F that is suitable for gradient descent techniques:

$$\frac{dx}{dt} = -\nabla_x E$$

With the notations of Sec.3.3, it leads to the equation

$$\frac{dx}{dt} = -BA^t \left[h - \tilde{F}(x)\right],$$

where B is an $m \times n$ matrix of elements

$$b_{kj} = \frac{x_k - z_{jk}}{\sigma_j^2} r_j(||x - z_j||),$$

and x_k and z_{jk} are the kth components of x and z_j. Since the model is analytically known, many second-order schemes are available (see [3]) and may result in faster convergence.

We show an example of such an "inversion" process in Fig. 6. The target is a random cell retina generated by a directional antenna. Initial conditions of parameters x are random. Subsequent images are generated by parameters x sampled from the inversion process. Sometimes the final estimated parameters do not generate the correct cell and the inversion gets stuck in some local minimum. In such cases, the solution is often to restart the inversion from other initial conditions. Those results have been obtained with 20 RBF, 30000 learning iterations and 1000 inversion iterations. RBF neural networks are site-dependant (in contrast to the convolutional neural network), which makes the learning time quite important. However, some generalization capacity has been observed and allows to start the training from a "generic" inital state in order to decrease the total learning time.

In the case of omnidirectional antennae, the function F to learn has only one parameter and RBF neural networks seem too heavy for the purpose of finding a best matching cell: very simple schemes such as the bissection method may suffice indeed.

In the case of many directional antennae on a site, an arbitrary initial solution is chosen, such as two antennae at 180° or three antennae at 120° and medium tilt and power. Then each antenna is configured on the basis of the cover hole it leaves when it is removed.

3.5 Enforcing traffic constraints

We recall that the traffic covered by an actual cell in the radio network must be lower than 43 Erlangs. The idea is to apply the same process as in cell inversion. First, an RBF neural network is trained to learn the relation: parameters $\mapsto$ traffic. If, after cell inversion, traffic exceeds 43 Erlangs, a backpropagation to the inputs (antenna parameters) causes the estimated traffic to converge towards 43 Erlangs. The drawback is that it may loose what the cell inversion had done. Both objectives (traffic and cover) are very often opposite and it is not clear whether the way traffic is corrected (priority given to power or azimuth) will be compatible with the main objective (cover).

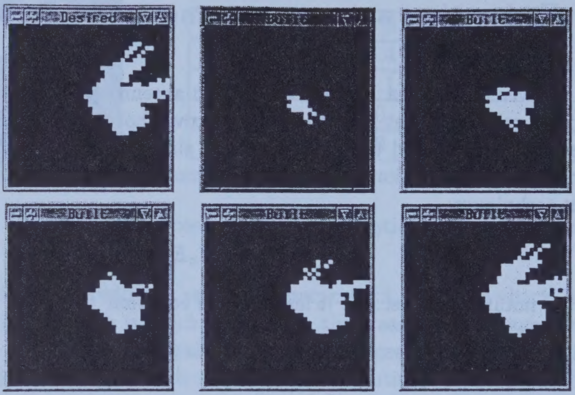


Figure 6: Inversion sequence. The first image is the target. Subsequent images are sampled from the inversion process.

4 A stochastic optimization algorithm

4.1 An analogy with statistical physics: interacting particle model

We have described a method for estimating antenna parameters but we still have to select a subset of potential sites and to decide for each of them how many antennae to put on it. For each site, the alternative reduces to:

- no antenna: the site is empty;
- one omnidirectional antenna;
- one directional antenna;
- two directional antennae;
- three directional antennae.

The choice of a site configuration leads (in a deterministic way) to antenna parameters and to a field strength distribution over the area. Sites are interacting through those overlapping field strength distributions: the shape of a cell reduces the free space available to other cells. This led us to consider sites as interacting particles taking discrete states (the continuous part has been studied in Sec. 3). The framework for studying large amounts of interacting particles is statistical physics. This framework has proved to be of great interest in the fields of image analysis and neural networks [4, 5].

We now refer to [4] and introduce the formalism of Markov random fields used in image restoration. Let Λ be the set of possible site configurations ($|\Lambda| = 5$) as described in the first paragraph and S denotes the finite set of potential sites. $\Omega = \Lambda^S$ is the set of all configurations over the whole radio network. Let $\omega = (x_s \in \Lambda, s \in S)$ be a configuration of all sites and $U(\omega)$ its cost as defined in Sec. 2.3, once antenna parameters have been estimated. Following the theory of Gibbs potential, we define a probability distribution π as

$$\pi(\omega) = \frac{1}{Z} e^{-U(\omega)/T},$$

where Z is a normalization constant (also called the partition function). T is a temperature as introduced by [6] and allows an annealing process during the algorithm. In the Gibbs potential theory, the cost function $U(\omega)$ is decomposed as $U(\omega) = \sum_{C \in \mathcal{C}} V_C(\omega)$, where $\mathcal{C}$ is the set of "cliques". A clique is a subset of sites S such that there exists an "interaction" between any pair of its elements. $V_C(\omega)$ only depends on configurations x_s of sites s in C. This property is at the heart of bayesian image restoration algorithms since it gives rise to great simplifications in the computation of conditional probability distributions.

The definition of cliques is straightforward for regular lattices in image processing but there is no such direct concept in radio networks. However there is a concept of electromagnetic range as modeled by the path loss (see Sec. 2.2).

4.2 Stochastic updating of site configurations

The algorithm builds a Markov chain over Ω: at each step, a site s is randomly selected and updated, i.e., a configuration for the selected site is sampled from the density $p_s(x)$. Let us explain how $p_s(x)$ is computed. Let ω be the current configuration and ω^x denotes the configuration ω where x_s at site s has been replaced by x. Then $p_s(x) = \pi(\omega^x)/\sum_{y \in \Lambda} \pi(\omega^y)$. For regular lattices, this expression simplifies and only involves cliques C such that site $s \in C$. Since cliques are not explicitly known in the case of radio networks, we have to evaluate the cost $U(\omega)$ of all possible configurations x for site s in order to compute $p_s(x)$ and then sample a configuration from it.

One notices that the algorithm updates sites *asynchronously*, much in the way Ising spin models or neural networks are simulated. There is no evidence of the existence of phenomenon such as "critical temperature" as observed and predicted for Ising spin models.

We claimed in the first paragraph that sites were selected randomly for updating. Actually, they may be selected from three different distributions, depending on the objective which may change during the algorithm:

- traffic distribution: high traffic regions should be served first;
- cover hole distribution: once high traffic regions have been covered, we turn to the remaining regions (with lower traffic demands but still to be covered);
- interference distribution: it is intended to localize "hard" regions.

5 Conclusion

We have presented new algorithms for the design and optimization of radio networks. Those algorithms rely on RBF neural networks for the continous part of the optimization, and on stochastic modeling, for the combinatorial part. RBF neural networks learn the relation between antenna parameters and reduced cells. Given a cover hole around a site, backpropagation to RBF inputs gives an estimation of antenna parameters whose cell best matches the cover hole. Reduced cells are computed by a convolutional neural network introduced in this paper. Finally, we have built a random field over the set of potential sites taking its values in the finite set of possible site configurations.

We have emphasized on an analogy with statistical physics and described sites as interacting particles: this analogy is fruitful in that it gives a good insight into the complex relationship (spatial, electromagnetic) between sites, without which no global optimization can be reached.

The optimization system has been developped in C++ language. Each of its components has been tested separately. Future work is well defined and includes full simulations of the model on real datasets of complexity ranging from 10 to 250 sites. Effort will especially bear upon the study of collective properties of the system and its optimization capacity.

References

[1] R. Linsker, "Self-organization in a perceptual network", *IEEE Computer*, vol. 21, no. 3, pp. 105–117, March 1988;

[2] T. Poggio and F. Girosi, "Networks for approximation and learning", *Proc. IEEE*, vol. 78, no. 9, pp. 1481–1497, September 1990;

[3] W. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical recipes in C*, second edition, Cambridge University Press, 1992;

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, November 1984;

[5] D.J. Amit, H. Gutfreund and H. Sompolinsky, "Spin-glass models of neural networks", *Physical Review A*, vol. 32, no. 2, pp. 1007–1018, August 1985;

[6] S. Kirkpatrick, C.D. Gelatt Jr. and M.P. Vecchi, "Optimization by simulated annealing", *Science*, 220, 1983.

A Short Review of Degree Theory on Wiener Space

A. Süleyman Üstünel
ENST, Dépt. Réseaux
46 Rue Barrault
75013 Paris, France
ustunel@res.enst.fr

Moshe Zakai
Department of Electrical Engineering
Technion—Israel Institute of Technology
Haifa 32000, Israel
zakai@ee.technion.ac.il

Let T be a C^1 map from $\mathbb{R}^n$ to $\mathbb{R}^n$ and proper (i.e. the inverse image of compact sets is compact). The degree theorem states that the degree q defined a.e. by

$$q = \sum_{y \in T^{-1}(x)} \text{sign det} \left(\frac{\partial T_i(y)}{\partial y_j} \right)_{n\times n} \tag{1}$$

does not depend on x and moreover for any bounded real valued function $\varphi(x), x \in \mathbb{R}^n$ with compact support,

$$\int_{\mathbb{R}^n} J(x)\varphi(Tx)dx = q \int_{\mathbb{R}^n} \varphi(x)dx \tag{2}$$

where $J(x)$ is the Jacobian of T: $\det\left(\frac{\partial T_i(x)}{dx_j}\right)_{n\times n}$. The notion of degree was extended in several directions and in particular, applied to establishing the existence of a solution x to non linear equations. In 1934 the notion of degree was extended by Leray and Schauder to a class of transformations on Banach space and applied to the proof of existence of solutions to certain partial differential equations [2], [3].

In 1986 E. Getzler [4] introduced the notion of degree for the shift transformations $Tw = w + u(w)$ on Wiener space, where u is an H-valued random variable. The results were improved in [5] and [6]. The results derived in these papers were extensions of (1) and (2) under strong integrability assumptions.

The following is an improvement to the results of [8] and [9].

Theorem A. *Let (W, H, μ) be an abstract Wiener space and let $u(w)$ be an H-valued random variable, $Tw = w + u(w)$. Assume that*

1. *for a.a. $w, h \mapsto u(w + h)$ is a continuous compact map on H.*
2. *for almost all w, for any $h_0 \in H$*

$$\sup\left\{|h| : h_0 = h + u(w + h)\right\} < \infty$$

μ-almost surely. (This condition is satisfied if, e.g., $\underline{\lim}|h_n + u(w + h_n)|_H \to \infty$ whenever $|h_n| \to \infty$.)

Let D_n *denote an increasing sequence of bounded subsets of* $H, D_n \nearrow H$ *as* $n \to \infty$. *Then, as* n *goes to* $+\infty$,

$$\deg\Big(T(w+h) - w, D_n, h_0\Big) \to q$$

almost surely, where q *is a non random constant. If* $q \neq 0$, *then* $\{\nu : T(\nu) = w\}$ *is a.s. non empty and the equation for* $\nu : \nu + u(\nu) = w$ *has a measurable solution* $\nu(w)$.

If (2)) is replaced by (2)':
(2)' for a.a. $w, |u(w+h)|_H = o(|h|_H)$ *as* $|h|_H \to \infty$,
then (2) is also satisfied and $q = 1$.

Note that in conditions (2) and (2)' the negligible sets on which the finiteness conditions fail may depend on h_0.

Combining Theorem A with a Sard inequality [9] and previously known results yields:

Theorem B. *Assume that* $u : W \to H$ *is* $H - C$ *and* $H - C^1_{\text{loc}}$ *with* $Q + H \subset Q$. *Suppose also that* $\sup(|h|_H : h + u(w+h) = h_0) < \infty$ *almost surely for any given* $h_0 \in H$ *and suppose also that the degree of* T *corresponding to* u *is nonzero, then* μ *is absolutely continuous with respect to* $T^*\mu$. *Denote by* S *the measurable right inverse of* T *whose existence was proved in the previous theorem. Then* μ *is absolutely continuous with respect to* $T^*\mu$. *Letting* L *be the corresponding Radon-Nikodym derivative, we have the following Girsanov-type identity:*

$$E\Big[f \circ T L \circ T\Big] = E[f]$$

for any $f \in C_b^+(W)$. *Moreover the following hold:*

i)

$$\mu(A) \leq \int \mathbf{1}_A \circ T |\Lambda_u| d\mu \,,$$

in particular $\mu \ll T^*\mu|_M$ *(hence* $\mu \approx T^*\mu|_M$*) with*

$$\frac{d\mu}{dT^*\mu|_M}(w) = \left[\sum_{y \in T^{-1}\{w\} \cap Q} \frac{1}{|\Lambda_u(y)|}\right]^{-1}.$$

ii)

$$u(S^{-1}(A)) \leq \int \mathbf{1}_A |\Lambda_A| \Lambda_u d\mu \,,$$

for an $A \in \mathcal{B}(W)$. *In particular* $S^*\mu \ll \mu|_M$ *and we have*

$$\sum_{y \in T^{-1}\{w\} \cap Q} \frac{1}{|\Lambda_u(y)|} \frac{dS^*\mu}{d\mu|_M}(y) = 1$$

for almost all $w \in W$.

As an application to a class of SPDEs, consider first the following corollary to Theorem A:

Corollary A. *Let K be a linear Hilbert-Schmidt operator on H and $G(h)$ a continuous function from H to itself such that $|G(h)|_H = o(|h|_H)$ as $|h|_H \to \infty$. Then the equation for y:*

$$y + KG(y) = \tilde{K}w \tag{3}$$

where $\tilde{K}$ is the natural extension of K to the abstract Wiener space, possesses a measurable H-valued solution.

Proof: Consider the shift:

$$Tw = w + G(\tilde{K}w)\,. \tag{4}$$

Since K is compact, $G(\tilde{K}w)$ satisfies the requirements on $u(w)$ in Theorem A. Therefore the equation $T\nu = w$ has a measurable solution, say $\nu = T^{-1}w$. Replacing w with $T^{-1}w$ in (4) and operating with $\tilde{K}$ yields

$$\begin{aligned}\tilde{K}w &= \tilde{K}\big(T^{-1}w + G(\tilde{K}T^{-1}w)\big)\\ &= \tilde{K}T^{-1}w + KG(\tilde{K}T^{-1}w)\end{aligned}$$

comparing with (3) yields that $y = \tilde{K}T^{-1}w$ solves (3). □

Let $\mathcal{D}$ be a bounded domain in $\mathbb{R}^d$, $\dot{w}$ a white noise on R^d and H_0 will denote the Hilbert space of real valued functions on $L^2(\mathcal{D})$. Let $g(x,r), x \in \mathcal{D}$, be real valued and such that for any $f \in H_0, g(x, f(x)), x \in R^d$ is a continuous and bounded transformation from H_0 to itself satisfying the assumption on G imposed in Corollary A with H replaced by H_0. We want to consider the stochastic partial differential equation

$$-\Delta\xi(x) + g\big(x, \xi(x)\big) = \dot{w}, x \in \mathcal{D} \tag{5}$$

where Δ is the Laplace operator on $\mathcal{D}$ with the boundary condition

$$\xi|_{\partial\mathcal{D}} = 0\,. \tag{6}$$

We restrict d to be 1,2, or 3 since, as shown by Buckdahn and Pardoux [1], $K = (-\Delta)^{-1}$ subject to the boundary condition (6) is a strictly positive Hilbert-Schmidt kernel and this is needed later. Equations (5), (6) can, therefore, be written as

$$\xi(x) + Kg\big(x, \xi(x)\big) = K\dot{w}\,. \tag{7}$$

Let R^d_x denote $\{y : y_i \le x_i, \forall i \le d\}$. Define the Cameron-Martin space H induced by H_0 as follows:

$$\eta(x) = \int_{R^d_x \cap \mathcal{D}} \xi(y)dy, \quad (\eta_1, \eta_2)_H = (\xi_1, \xi_2)_{H_0}\,.$$

Set

$$\big(G_1(\eta)\big)(x) = \int_{R^d_x \cap \mathcal{D}} g\big(y, \xi(y)\big)dy\,,$$

and

$$(K_1\eta)(x) = \int_{R^d_x \cap \mathcal{D}} \int_{\mathcal{D}} K(y_1, y_2)\xi(y_2)dy_2dy_1\,.$$

Then equation (7) is equivalent to:

$$\eta + K_1G_1(\eta) = \tilde{K}_1w \tag{8}$$

and the existence of a measurable solution to (8) follows from Corollary A. □

References

[1] R. Buckdahn and E. Pardoux, "Monotonicity methods for white noise driven quasilinear SPDEs". In *Progress in Probability*, M. Pinsky (Editor), vol. 2, Birkhauser, 1993.

[2] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, 1985.

[3] I. Fonseca and W. Gangbo, *Degree Theory in Analysis and Applications.* Oxford Lecture Series in Mathematics and Its Applications 2. Clarendon Press, Oxford, 1955.

[4] E. Getzler, "Degree theory for Wiener maps", *J. Funct. Anal.*, vol. 68, pp. 388–403, 1986.

[5] S. Kusuoka, "Some remarks on Getzler's degree theorem". In *Proc. of 5th Japan USSR Symp.* (Lecture Notes Math, vol. 1299, pp.239–249), Berlin Heidelberg New York: Springer 1988.

[6] A.S. Üstünel and M. Zakai, "Applications of the degree theorem to absolute continuity on Wiener space", *Prob. Theory Rel. Fields*, vol. 95, pp. 509–520, 1993.

[7] A.S. Üstünel and M. Zakai, "Transformation of the Wiener measure under non-invertible shifts", *Prob. Theory Rel. Fields*, vol. 99, pp. 485–500, 1994.

[8] A.S. Üstünel and M. Zakai, "Degree theory of Wiener space", *Prob. Theory Rel. Fields*, vol. 108, pp. 259–279, 1997.

[9] A.S. Üstünel and M. Zakai, "The Sard inequality on Wiener space", *J. Functional Analysis*, vol. 149, pp. 226–244, 1997.

Stochastic Control and Linear Programming

Larry Shepp

Rutgers U. Statistics Dep't, Hill Center, Room 569

New Brunswick, NJ, 08855-1179

e-mail: shepp@stat.rutgers.edu

Abstract. I will discuss several optimal stochastic control problems which can be solved in explicit form by the same technique. In each case one first guesses the answer, i.e., the optimal control and the value at the optimum, by various heuristics, in each case employing some intuition about the specific problem. Once the answer is successfully guessed, the martingale apparatus is used in a crank-turning way to prove that the guessed answer is correct.

This is formally quite similar to the way that game theory and general infinite dimensional linear programming problems are solved, where an extreme point x of a convex set C is sought which maximizes a linear functional $x \cdot y$ over $x \in C$. Here too one solves the problem by guessing by various heuristics the extreme point x_0 and the value m of the maximum. Then one shows that the plane $\pi : x \cdot y = m$ through x_0 is a support plane for C. Now the proof is achieved in a crank-turning way: one verifies that the inequality $x \cdot y \leq m$ holds for all points $x \in C$. This proves the maximum cannot be any larger than m and the fact that x_0 is on π completes the proof.

There is a good reason for this similarity: this class of stochastic optimal control problems is a special case of linear programming. This observation does not seem to be well known. Since many stochastic control problems problems have already been solved without everybody knowing this analogy, one could argue that it may not be so important, but this is a different matter. It does seem insightful to realize it. I will try to convince you that this analogy is true and interesting and also that the problems solved in this way can give insight into the real world.

Section 1. Some stochastic control problems with simple solutions.

The class of optimal control problems where the solution can be explicitly obtained is rather special. The problem must be linear and apparently also one-dimensional in some sense. The surprise is these restrictions, although stringent, do not exclude many problems of interest. I will first give some examples.

Finance models.

Consider a firm faced with the need to define itself; it has n corporate options, each involving risk σ_i and each having its own potential μ_i for generating profit. One can think of the different options as n different products which the firm can produce, and any given time, a convex combination of products can be obtained by rapid switching, since there are no costs in this simple model for switching between product lines[1]. Let $I_t \in \{1, \ldots, n\}$ be the actual product produced at time t.

The firm extracts profit by paying dividends and Z_t, $t \geq 0$ is the total dividend up to time t. We suppose the dividend $dZ_t \geq 0$ paid in dt reduces the capital stream X_t of the firm, which evolves according to the Bachelier model, $X_0 = x$, $dX_t = \mu_{I_t} dt + \sigma_{I_t} dW_t - dZ_t$, $0 \leq t \leq \tau_0$ where W_t is a Brownian motion or Wiener process and $I_t \in \{1, 2, \ldots, n\}$ at each time t is the actual corporate option used at time t. Let $A = \{(\mu_i, \sigma_i^2), i = 1, \ldots, n\}$ be the set of drift-volatility pairs as a set of points in the plane. Note that the firm goes bankrupt at the first time $t = \tau_0 \leq \infty$ at which $X_t = 0$ and then $dX_t = dZ_t \equiv 0$ for $t \geq \tau_0$. It is reasonable to assume that the firm attempts to maximize $E \int_0^\infty e^{-rt} dZ_t$ the expected total discounted future value of its dividends by choosing both a strategy I_t and a dividend stream Z_t so as to maximize $V(A, x) = V(x) = E_x \int_0^\infty e^{-rt} dZ_t$. The processes I_t and Z_t must be nonclairvoyant, i.e., they cannot use future behavior of W_u, $u > t$. This is a typical stochastic control problem with I_t and Z_t as the control variables. It may not appear to be a problem of linear programming, but it is.

If there are a pair of points $(\mu_1, \sigma_1) \in A$ and $(\mu_2, \sigma_2) \in A$ with $\mu_1 < \mu_2$ and $\sigma_2^2 \leq \sigma_1^2$ or even if $\sigma_2^2/\mu_2 \leq \sigma_1^2/\mu_1$ then one should eliminate (μ_1, σ_1) from A, i.e., $V(A - \{(\mu_1, \sigma_1)\}, x) \equiv V(A, x)$. In other words, (μ_1, σ_1) is uniformly dominated by (μ_2, σ_2), an option of greater profit-potential and smaller risk independent of the other points of A and independent of the discount rate r. In order that each pair be used at some point it is necessary that the points of A be strictly monotonic and concave[RS]. In this case we can give the optimal strategy and the value $V(A, x)$ of the firm explicitly. The formula for $V(A, x)$ is not completely trivial even if A is a singleton.

It is no loss of generality to assume that $\mu_1 < \mu_2 < \ldots < \mu_n$, and then there exist thresholds $0 = a_0 < a_1 < a_2 < \ldots < a_n$ such that we use the j^{th} most conservative strategy (μ_j, σ_j) when $a_{j-1} < x < a_j$. Perhaps not surprisingly, the firm should pay dividends *only* when its fortune X_t is sufficiently large, $X_t > a$ and the value for the threshold a is found. We guess $V(x) = \hat{V}(x)$ by looking for V to be in C^2 (the smoothness heuristic[2]) and noting that if the above form is optimal then inside the interval $[a_{j-1}, a_j]$ we expect V to satisfy the second order differential equation

---

[1] Later we will consider a similar problem with switching costs

$\frac{1}{2}\sigma_j^2 V''(x) + \mu_j V'(x) = 0, \text{ for } x \in [a_{j-1}, a_j]$

because the policy doesn't change. There are thus $2n$ constants to solve for and n values of a for a total of $3n$ unknowns. But the condition that V ought to be in C^2 gives the right number of conditions to determine all the unknowns and enable us to guess $\hat{V}$. For details see[RS].

We have made strong use of exponential discounting; the problem would be not one-dimensional and would be really impossible to solve explicitly if we had to keep track of time as well as x. This exponential discounting serves to get rid of time and reduces the problem to deciding what to do when the capital on hand is x, for each x.

The proof is as usual by martingale methods, provided we know the answer. If one doesn't know the answer in explicit form it is hard to see how anything can be proven rigorously. Direct proofs as in[Sh69] are simply impossible and wrong. The reason it is necessary to use martingales is that the condition of non-anticipation is impossible to take account of without martingale theory. Of course if the strategy is allowed to anticipate the future increments of W, one could get strictly greater profit.

To prove the result, consider the process

$Y_t = e^{-rt}\hat{V}(X_t) + \int_0^t dZ_u$

which formally represents the value at time t under any corporate strategy which the firm may employ, suboptimal or not. It's easy to see that Y_t is always a super-martigale, i.e., expectation decreasing. Thus for any stopping time τ, say the time of bankruptcy,

$EY_\tau \leq Y_0$

But $Y_0 = \hat{V}(x)$ and $Y_tau = \int_0^\tau dZ_u$, so this inequality says that $V(x) \leq \hat{V}(x)$ because X_t can be arbitrary. Thus one proves that one can do no better than the guessed value and one then shows easily that if the optimal strategy is followed, both in the choice of policy at each time, and in the dividend rule, then Y_t is a uniformly integrable martingale and equality is achieved, i.e., $V(x) \geq \hat{V}(x)$. It is to be noted that dividends are paid only when $X_t = a_n$ (in the so-called local time at this boundary) and this keeps X_t in the compact interval $[0, a_n]$.

---

[2] In guessing the answer it is often necessary to use the well-known principle, due to Herman Chernoff[Cher, BaCh] and A N Kolmogorov, that the optimum tends to be a bit smoother than its competitors.

Since its stochastic oscillations are bounded away from 0, it is easy to see that $\tau < \infty$ with probability 1, i.e., the company goes bankrupt in a finite time almost surely. This seems a little surprising at first, but if the company aggressively seeks to maximize profits it will be like real firms which walk on the edge and go bankrupt eventually.

Where is the linear programming? It's not so obvious where it is. In a formal sense, the free variables of the linear program must keep track of the information structure so that the nonclairvoyance of the strategy is encoded. This is possible because the condition that r.v.'s are independent is a linear condition because X and Y are independent if and only if the linear functionals $Ef(X)g(Y) = Ef(X)Eg(Y)$ holds for all bounded f and g. The unknowns of the linear program represent the control variables I and Z but the space in which all this is happening is very abstract, especially in continuous time. We may write the linear constraints, $0 \le I_i(t) \le 1$ and $\sum_{i=1}^{n} I_i(t) = 1$, so that $I_i(t)$ represents the fraction of time that option i is being used at time t. Similarly, $dZ_t \ge 0$ is a linear constraint and the nonclairvoyance conditions on both I_i and Z are linear. The objective function becomes a linear form in the unknowns. The role of the supporting hyperplane is played by the supermartingale (submartingale if we are minimizing rather than maximizing). I have tried to convince you that linear programming is involved in stochastic control via martingales without giving a complete proof. I hope that examples will make it clearer.

Stochastic control problems which are tractable are few and far between, just as not all linear programming problems have explicit solutions, but if there is more continuity, the problem often becomes easier because calculus, especially Ito calculus, involves differential rather than difference equations and the optimal strategy may sometimes be guessed. In discrete case explicit solutions are rarer and rigorous proofs of optimality are consequently more difficult[CRS]. But more than smoothness is needed to get explicit answers: in all cases I know when explicit solution is possible, there is something one-dimensional going on. This one-dimensionality may be rather subtle however. I'll give some examples of this. I have discussed some aspects of the following example earlier in[S82].

Optimal stochastic control of a satellite[BSW].

A communications satellite is kept in geo-stationary orbit by controlling its position coordinates with a rocket which has a given quantity Y_t of fuel at each time t. Suppose the value of the x coordinate at time t is X_t (one could consider a higher dimensional problem, but let's keep it simple) at time t is X_t, and that $X_0 = x$, $Y_0 = y$, $dY_t = -|u|dt$ and

$dX_t = udt + \sigma dW_t, 0 \leq t$. We can control X_t (until we run out of fuel) by expending fuel, $u = u_t$, but we must keep $\int_0^\infty |u_t|dt \leq y$, where y is the total fuel on board at time $t = 0$. Suppose we want to make the position X_t as close to zero as possible and prefer to make errors later rather than sooner. We might model the problem as one of minimizing the total discounted mean square position. That is, we want to minimize

$E \int_0^\infty e^{-rt} {X_t}^2 dt,$

by choice of the control, u.

It's quite intuitive based on the Bellman equation[BSW] that the optimal control rule ought to be of the form: "expend no fuel if $|X_t| \leq a(Y_t)$", where Y_t is the amount of fuel left at time t. If $X_t > a(Y_t)$, then u should be taken infinitely negative to instantaneously (bang-bang control) drive X back towards zero, while if $X_t < -a(Y_t)$, then u should be taken large and positive to again make X_t move towards zero. Of course when we run out of fuel, then W takes over, and we can do nothing more. The free boundary, $a(y)$ can be found explicitly[BSW], and is a completely explicit curve which the reader may enjoy trying to discover.

It appears that the problem is two-dimensional since we are looking for the value $V(x, y) = min_u E_{x,y} \int_0^\infty e^{-rt} {X_t}^2 dt$, but if this were the case then we probably couldn't solve the problem in explicit form. Since the answer involves a curve, it also appears to be two-dimensional. It isn't and the reason is two-fold. For one, when $|X_t| < a(Y_t)$ the motion of X is one-dimensional and a threshold rule is needed for when to apply fuel as long as $y = Y_t$ is unchanging, while we are letting the satellite drift. In this case $V(x, y)$ satisfies an ordinary differential equation in the region $|x| < a(y)$. Secondly, if the problem were genuinely two dimensional, then V would satisfy a partial differential equation and no explicit solution could be found (at least I know of no such examples which cannot be reduced to one-dimension in a trivial way). If the answer cannot be written down in a completely explicit way, the martingale argument below could hardly be carried out. Note that existence theorems seem to be of little or no value in such problems because one needs to check the martingale and super-martingale conditions to get a proof. This is not to say that numerical solutions are impossible, but a rigorous proof requires martingales which must be explicit, it seems.

The martingale argument is quite similar to the first example. If X_t is the position at time t for any choice of fuel expenditure (non-clairvoyant!), and if we could write down a correct guess at the optimal value $\hat{V}(x, y)$ of $V(x, y)$, then $Z_t = e^{-rt}\hat{V}(X_t, Y_t) + \int_0^t X_s^2 e^{-rs} ds$

would be a local *sub*-martingale and we easily obtain $V(x,y) \geq \hat{V}(x,y)$. Again under optimal stochastic control, the process Z_t is a martingale.

How does one guess the exact value of $V(x,y)$? We suspect that $V(x,y)$ satisfies the ordinary differential equation,

$$\tfrac{1}{2}V''(x,y) - rV(x,y) + x^2 = 0$$

where V'' is the second derivative with respect to x inside the region $|x| < a(Y)$, where we still don't know the function $a = a(y)$. This at least is a ordinary differential equation, and there are two constants which are also functions of y to be determined. Smoothness now comes to the rescue: we want $V_x(x,y) = \pm V_y(x,y)$ (V_x and V_y denote partial derivatives) along the curve $x = \pm a(y)$ because this differential relation on V holds *outside* the curve and so must also hold *on* the curve as well. It holds outside the curve because outside the curve we presume intuitively that fuel is expended instantaneously and since each decrease in y involves a corresponding and equal increase or decrease in x the differential relation must hold. We want V to be twice differentiable in x so that Ito calculus can be used and this leads to a three parameter family of equations which can be solved to determine both the curve a and the two constant functions of y in the differential equation in x on each line $y =$ constant. For more details see[BSW]. In small variations on the problem such as a finite rate of fuel expenditure, or non-exponential discounting, explicit answers become impossible and one must resort to numerics. Note that again we are dealing with a linear programming problem where the martingale argument is playing the role of the supporting hyperplane. I solved both these problems and others below before I realized the close analogy to linear programming.

Hiring and Firing.

Consider the following simple stochastic model for determining the optimal policy for the size of staff of a large (especially, but not only, research) corporation as a function of its current capital. The objective is to provide a quantitative theory with explicit prescriptions for increasing, decreasing or leaving constant the number of personnel in order to maximize total profit, discounted with fixed interest rate. This model and problem is close in spirit to that of the first problem above except that it has a new feature, switching costs. It costs the company to increase or to decrease its staff size.

Features of the optimal policy seem in at least qualitative agreement with observed practice in that, for example, the optimal policy includes large scale down-sizings similar to large layoffs or incentive retirement plans observed in real corporations. In the optimal plan, this type of firing occurs *only on an upswing*, i.e. when the company's fortune *ascends* to a threshold relative to staff size.

Small scale or "attrition" firing also occurs under the optimal policy, but this type of firing occurs *only on a downswing*, when the company's fortune *descends* to a threshold relative to staff size. The model is extremely simple and depends on only 5 estimable parameters; thus it should be useful for guiding, or at the least stimulating, thought about corporate sizing policy.

Defining the model

Suppose the fortune, or available capital, of a research corporation at time t is X_t, $t \geq 0$, where $x_0 = x$ and for $0 < t < \tau$

$$dX_t = -dZ_t + U_t(\mu dt + \sigma dW_t) - k_- d_- U_t - k_+ d_+ U_t \tag{1}$$

describes the incremental evolution of the fortune, where W is a standard Wiener process, U_t is the level of research or the "size" of the (research) staff at time t, which is under our control and may fluctuate up or down with $U_0 = u$, and $U_t \geq 0$ for all t. Z_t, the total profit or dividend extracted up to time t ($Z_0 = 0$) is assumed, of course, nondecreasing, $dZ_t \geq 0$, and to also be under our control.

We want to select U_t and Z_t to *maximize* the time-discounted *total profit*,

$$V(x, u) = \sup E_{x,u} \int_0^\tau e^{-rt} dZ_t \tag{2}$$

where τ represents the time of bankruptcy: $\tau = t$ if $X_t = 0$ for the first time, $0 \leq \tau \leq \infty$. The supremum in 2 is taken over all strategies, i.e. choices of the controls, U_t and Z_t, which are arbitrary except that $dZ_t \geq 0$, $U_t \geq 0$, and each is *nonclairvoyant* so that *future* increments of W are independent of U_t and Z_t at time t. The interest rate, $r > 0$, is assumed known and constant.

Increments of U_t which are positive (hiring) are denoted d_+U_t and those which are negative (firing) by $-d_-U_t$ so that $dU(t) = d_+U_t - d_-U_t$ with $d_+U_t \geq 0$, $d_-U_t \geq 0$. We assume that the *one-time cost* of increasing U by a unit (employee) is a known constant k_+ which we assume nonnegative. Similarly, the one time cost of decreasing U by a unit is a known constant k_-, also here assumed nonnegative. Note k_+ and k_- *do not include* salary or maintenance costs, but instead represent *one-time* costs involved in hiring (relocation, etc.) or firing (severance, etc.). A larger U_t in 1 makes for greater ruin exposure, as we'll see. It is easy to see that unless $k_+ = k_- = 0$ (in which case U_t is irrelevant and the problem reduces to [RS]), wlog U_t in 1 can be assumed of bounded variation so that d_-U_t and d_+U_t are well-defined.

In 1, μ represents the *mean* income per unit of research investment per unit time *net of*, (i.e., *above*) *salary and maintenance.* Of course μ depends on the intrinsic *quality* of the staff, but this is assumed to be held constant through replacement of older or worn equipment and personnel. This process is called maintenance and μ represents mean income above this maintenance (and above overhead and salary). One can show[SS4] show that unless $\mu > 0$, there is no value in the research business under this model; namely if $\mu \leq 0$ then we show $V(x, u) \equiv x$ in 2: the company should distribute all of its capital and go bankrupt at once. This policy is called, "take the money and run".

Of course success in research is not predictable, so a random noise term is particularly appropriate if the corporation is in research and in 1 this is modelled by adding σdW_t to μdt. Note that, in the real world, σ includes other uncertainties, e.g. the uncertainty of the market value of the products of the company; indeed $\sigma dW_t + \mu d_t < 0$ is possible since research can produce negative effects (e.g., product liability, etc.).

The model given depends on only 5 parameters, r, μ, σ, k_+, k_- each of which should be easily estimable based on past statistics of the firm. In fact there are really only 4 independent parameters because μ, r, and σ^2 are linked by the choice of the time unit. An inelegance is that we must assume that the size of the company is constrained to be bounded by a known quantity M. If the company can grow arbitrarily the maximum profit becomes infinite. Considerations of lack of efficiency when a firm grows too large which are not included in this simple model would make the M assumption unnecessary, I am sure.

The solution to this stochastic control problem is given in[SS4] and gives much insight into real hiring and firing: the company should not hire or fire when its capital is small and its staff size is large, but when and if its capital becomes large enough it should engage in massive layoffs. It is often observed that real firms layoff on an upturn in their economic performance which seems consistent with this theory, however, in this model, as discussed in[SS4] layoffs only happen once, after which the firm relies only on attrition firing under optimal management. Real firms often downsize more than once and so either there are some additional considerations, or the firms are not following optimal policy of maximizing profit. In any case, it seems surprising that such relevant conclusions can be obtained with such a simple model. For more discussion of the model see[SS4].

In this problem the optimal solution is rather hard to guess. There are actually 3 free boundaries of the form $x = f_i(u), i = 1, 2, 3$ to find. Here f_1 is the curve of values of x and u along which mass firing takes place as discussed above, which I call the "firing line"; f_2, along which attrition firing takes place; and f_3 along which hiring takes place. Each curve is a straight line, but guessing that this is the form of the optimum apparently needs some intuition about the real problem. The company pays dividends only when $u = M$, so the firm is at maximal size, and the capital $x \geq Q$ is also maximal. Again the phase space of the problem is compact and the diffusion coefficient is bounded below, and again the firm goes bankrupt under its aggressive optimal profit making strategy with probability one in finite time.

It seems amazing that such a simple model can have so much in common with reality and yet be new, but this seems to be the case. This work can be viewed as an extension to include switching costs of the problem a corporation faces in deciding between various available corporate policies, studied above[RS] which gives further discussion of bankruptcy. A referee of[SS4] asked that I include the following caveat which I suppose I am morally bound to repeat here: "Everybody understands that this work is an elegant mathematical essay, rather than a practical recommendation for firing and hiring".

In all the problems discussed here, there is one-dimensionality in some aspect if one looks closely, although each problem seems at first glance to have more than one free variable.

There are many other problems of this type that could be discussed including hedge option pricing[SS1, SS2, SS3, Sh69, DSS], but I have run into my 10 page limit and will quit here, hoping to have proven at least some of my points.

References

[B] Bachelier, L. (1900), Théorie de la Speculation. Reprinted in Cootner (ed.), (1967), The random character of stock market prices, MIT Press, Cambridge, Mass., 17–78.

[BSW] Benes, V. E., Shepp, L. A. and Witsenhausen, H. S. (1980). Some solvable stochastic control problems, Stochastics 4, 39–83.

[S82] Shepp, LA An explicit solution to a problem in nonlinear stochastic control, Nonlinear Stochastic Problems, ed. by R.S.Bucy and J.M.F.Moura D.Reidel Dordrecht, Holland vol 104C, 229-232, 1982

[Cher] Chernoff, Herman (1961). Sequential tests for the mean of a normal distribution. Proc. 4th Berk. Symp. Math. Stat. Prob. 1, 79–91, University of California Press.

[BaCh] Bather, John and Chernoff, Herman (1967). Sequential decisions in the control of a spaceship. Proc. 5th Berk. Symp. Math. Stat. Prob. 3, 181–207, University of California Press.

[Sh69] Shepp, L. A. (1969). Explicit solutions to some problems of optimal stopping, Ann. Math. Stat., vol. 40, 993–1010.

[CRS] Chow, Y. S., Robbins, Herbert, and Siegmund, David (1991). The Theory of Optimal Stopping, Dover Pubs.

[DSS] Dubins, L., Shepp, L. A., and Shiryaev, A. N., Optimal stopping rules and maximal inequalities for Bessel processes, Teor. Veroy. Prim. 38, 288-330.

[RS] Radner, R. and Shepp, L. Risk vs. Profit Potential: A Model for Corporate Strategy Journal of Economic Dynamics and Control 20, (1996), 1373-1393.

[SS1] Shepp, L. A. and Shiryaev, A. N. (1993). The Russian option: reduced regret, Ann. Appl. Prob. 3, 631-640.

[SS2] Shepp, L. A. and Shiryaev, A. N. A new look at the pricing of Russian options, Theory of Probab. and its Appl., 1993, N3, to appear.

[SS3] Shepp, L. A. and Shiryaev, A. N. A dual Russian option for selling short, Probability Theory and Mathematical Statistics, vol. ed. by Ibragimov and Zaitsev, Gordon and Breach (1996), 209-218

[SS4] Shepp, L. A. and Shiryaev, A. N. Hiring and firing optimally in a large corporation, Journal of Economic Dynamics and Control 20, (1996), 1523-1539.

Coding Theorems for "Turbo-Like" Codes*

Dariush Divsalar, Hui Jin, and Robert J. McEliece
Jet Propulsion Laboratory and California Institute of Technology
Pasadena, California USA
E-mail: `dariush@shannon.jpl.nasa.gov, (hui, rjm)@systems.caltech.edu`

Abstract.

In this paper we discuss AWGN coding theorems for ensembles of coding systems which are built from fixed convolutional codes interconnected with random interleavers. We call these systems "turbo-like" codes and they include as special cases both the classical turbo codes [1,2,3] and the serial concatentation of interleaved convolutional codes [4]. We offer a general conjecture about the behavior of the ensemble (maximum-likelihood decoder) word error probability as the word length approches infinity. We prove this conjecture for a simple class of rate $1/q$ *serially concatenated codes where the outer code is a* q*-fold repetition code and the inner code is a rate* 1 *convolutional code with transfer function* $1/(1+D)$*. We believe this represents the first rigorous proof of a coding theorem for turbo-like codes.*

1. Introduction.

The 1993 discovery of turbo codes by Berrou, Glavieux, and Thitimajshima [1] has revolutionized the field of error-correcting codes. In brief, turbo codes have enough randomness to achieve reliable communication at data rates near capacity, yet enough structure to allow practical encoding and decoding algorithms. This paper is an attempt to illuminate the first of these two attributes, i.e., the "near Shannon limit" capabilities of turbo-like codes on the AWGN channel.

Our specific goal is to prove AWGN coding theorems for a class of generalized concatenated convolutional coding systems with interleavers, which we call "turbo-like" codes. This class includes both parallel concatenated convolutional codes (classical turbo codes) [1, 2, 3] and serial concatenated convolutional codes [4] as special cases. Beginning with a code structure of this type, with fixed component codes and interconnection topology, we attempt to show that as the block length approaches infinity, the ensemble (over all possible interleavers) maximum likelihood error probability approaches zero if E_b/N_0 exceeds some threshold. Our proof technique is to derive an explicit expression for the ensemble input-output weight enumerator (IOWE) and then to use this expression, in combination with either the classical union bound, or the recent "improved" union bound of Viterbi and Viterbi [9], to show that the maximum likelihood word error probability approaches zero as $N \to \infty$. Unfortunately the difficulty of the first step, i.e., the computation of the ensemble IOWE, has kept us from full success, except for some very simple coding systems, which we call *repeat and accumulate* codes. Still, we are optimistic that this technique will yield coding theorems for a much wider class of interleaved concatenated codes. In any case, it is satisfying to have rigorously proved coding theorems for even a restricted class of turbo-like codes.

Here is an outline of the paper. In Section 2 we quickly review the classical union bound on maximum-likelihood word error probability for block codes on the AWGN

---

* Dariush Divsalar's work, and a portion of Robert McEliece's work, was performed at JPL under contract with NASA. The remainder of McEliece's work, and Hui Jin's work, was performed at Caltech and supported by NSF grant no. NCR-9505975, AFOSR grant no. 5F49620-97-1-0313, and grant from Qualcomm.

channel, which is seen to depend on the code's weight enumerator. In Section 3 we define the class of "turbo-like" codes, and give a formula for the average input-output weight enumerator for such a code. In Section 4 we state a conjecture (the interleaver gain exponent conjecture) about the ML decoder performance of turbo-like codes. In Section 5, we define a special class of turbo-like codes, the repeat-and-accumulate codes, and prove the IGE conjecture for them. Finally, in Section 6 we present performance curves for some RA codes, using an iterative, turbo-like, decoding algorithm. This performance is seen to be remarkably good, despite the simplicity of the codes and the suboptimality of the decoding algorithm.

2. Union Bounds on the Performance of Block Codes.

In this section we will review the classical union bound on the maximum-likelihood word error probability for block codes.

Consider a binary linear (n, k) block code C with code rate $r = k/n$. The *(output) weight enumerator* (WE) for C is the sequence of numbers $A_0, \ldots, A_n$, where A_h denotes the number of codewords in C with (output) weight h. The *input-output weight enumerator* (IOWE) for C is the array of numbers $A_{w,h}$, $w = 0, 1, \ldots, k$, $h = 0, 1, \ldots, n$: $A_{w,h}$ denotes the number of codewords in C with input weight w and output weight h.

The union bound on the word error probability P_W of the code C over a memoryless binary-input channel, using maximum likelihood decoding, has the well-known form

$$P_W \leq \sum_{h=1}^{n} A_h z^h \tag{2.1}$$

$$= \sum_{h=1}^{n} \left(\sum_{w=1}^{k} A_{w,h} \right) z^h. \tag{2.2}$$

In (2.1) and (2.2), the function z^h represents an upper bound on the pairwise error probability for two codewords separated by Hamming (output) distance h. For AWGN channels, $z = e^{-rE_b/N_0}$ where E_b/N_0 is the signal-to-noise ratio per bit.

3. The Class of "Turbo-Like" Codes.

In this section, we consider a general class of concatenated coding systems of the type depicted in Figure 1, with q encoders (circles) and $q - 1$ interleavers (boxes). The ith code C_i is an (n_i, N_i) linear block code, and the ith encoder is preceded by an interleaver (permuter) P_i of size N_i, except C_1 which is not preceded by an interleaver, but rather is connected to the input. The overall structure must have no loops, i.e., it must be a graph-theoretic tree. We call a code of this type a "turbo-like" code.

Define $s_q = \{1, 2, \ldots, q\}$ and subsets of s_q by $s_I = \{i \in s_q : C_i \text{ connected to input}\}$, $s_O = \{i \in s_q : C_i \text{ connected to output }\}$, and its complement $\bar{s}_O$. The overall system depicted in Figure 1 is then an encoder for an (n, N) block code with $n = \sum_{i \in s_O} n_i$.

If we know the IOWE $A^{(i)}_{w_i,h_i}$'s for the constituent codes C_i, we can calculate the *average* IOWE $A_{w,h}$ for the overall system (averaged over the set of all possible interleavers), using the *uniform interleaver* technique [2]. (A uniform interleaver is defined as a probabilistic device that maps a given input word of weight w into all distinct $\binom{N_i}{w}$ permutations of it with equal probability $p = 1/\binom{N_i}{w}$.) The result is

$$A_{w,h} = \sum_{\substack{h_i : i \in s_O \\ \Sigma h_i = h}} \sum_{h_i : i \in \bar{s}_O} A^{(1)}_{w_1,h_1} \prod_{i=2}^{q} \frac{A^{(i)}_{w_i,h_i}}{\binom{N_i}{w_i}} \tag{3.1}$$

In (3.1) we have $w_i = w$ if $i \in s_I$, and $w_i = h_j$ if C_i is preceeded by C_j (see Figure 2.). We do not give a proof of formula (3.1), but it is intuitively plausible if we note that the term $A^{(i)}_{w_i,h_i}/\binom{N_i}{w_i}$ is the probability that a random input word to C_i of weight w_i will produce an output word of weight h_i.

For example, for the $(n_2+n_3+n_4, N)$ encoder of Figure 1 the formula (3.1) becomes

$$
\begin{aligned}
A_{w,h} &= \sum_{\substack{h_1,h_2,h_3,h_4 \\ (h_2+h_3+h_4=h)}} A^{(1)}_{w_1,h_1} \frac{A^{(2)}_{w_2,h_2}}{\binom{N_2}{w_2}} \frac{A^{(3)}_{w_3,h_3}}{\binom{N_3}{w_3}} \frac{A^{(4)}_{w_4,h_4}}{\binom{N_4}{w_4}} \\
&= \sum_{\substack{h_1,h_2,h_3,h_4 \\ (h_2+h_3+h_4=h)}} A^{(1)}_{w,h_1} \frac{A^{(2)}_{w,h_2}}{\binom{N}{w}} \frac{A^{(3)}_{h_1,h_3}}{\binom{n_1}{h_1}} \frac{A^{(4)}_{h_1,h_4}}{\binom{n_1}{h_1}}.
\end{aligned}
$$

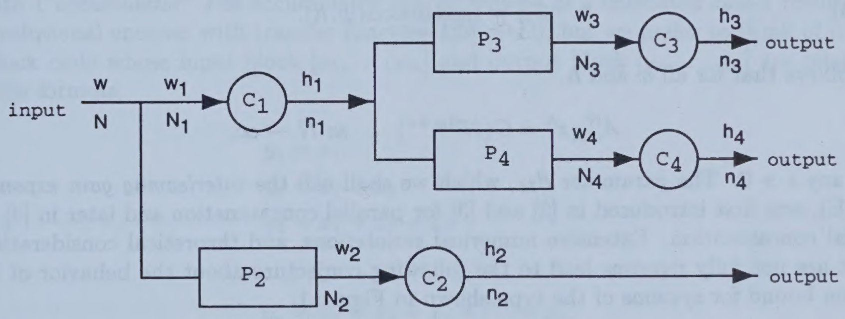


Figure 1. A "turbo-like" code with $s_I = \{1,2\}, s_O = \{2,3,4\}, \bar{s}_O = \{1\}$.

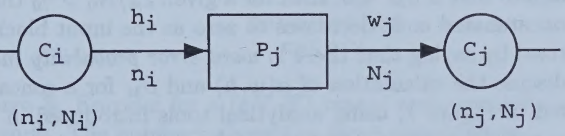


Figure 2. C_i (an (n_i, N_i) encoder) is connected to C_j (an (n_j, N_j) encoder) by an interleaver of size N_j. We have the "boundary conditions" $N_j = n_i$ and $w_j = h_i$.

4. The Interleaving Gain Exponent Conjecture.

In this section we will consider systems of the form depicted in Figure 1, in which the individual encoders are truncated convolutional encoders, and study the behavior of the average ML decoder error probability as the input block length N approaches infinity. If $A^N_{w,h}$ denotes the IOWE when the input block has length N, we introduce the following notation for the union bound (2.2) for systems of this type:

$$P_W^{\mathrm{UB}} \overset{\mathrm{def}}{=} \sum_{h=1}^{n} \left(\sum_{w=1}^{N} A^N_{w,h} \right) z^h. \tag{4.1}$$

Next we define, for each fixed $w \geq 1$ and $h \geq 1$,

$$\alpha(w,h) = \limsup_{N \to \infty} \log_N A^N_{w,h}. \tag{4.2}$$

It follows from this definition that if w and h are fixed,

$$A^N_{w,h} z^h = O(N^{\alpha(w,h)+\epsilon}) \qquad \text{as } N \to \infty,$$

for any $\epsilon > 0$. Thus if we define

$$\beta_M = \max_{h \geq 1} \max_{w \geq 1} \alpha(w,h). \tag{4.3}$$

it follows that for all w and h,

$$A^N_{w,h} z^h = O(N^{\beta_M+\epsilon}) \qquad \text{as } N \to \infty,$$

for any $\epsilon > 0$. The parameter β_M, which we shall call the *interleaving gain exponent* (IGE), was first introduced in [2] and [3] for parallel concatenation and later in [4] for serial concatenation. Extensive numerical simulations, and theoretical considerations that are not fully rigorous lead to the following conjecture about the behavior of the union bound for systems of the type shown in Figure 1.

The IGE Conjecture. *There exists a positive number γ_0, which depends on the q component convolutional codes and the tree structure of the overall system, but not on N, such that for any fixed $E_n/N_0 > \gamma_0$,as the block length N becomes large,*

$$P_W^{\mathrm{UB}} = O(N^{\beta_M}) \tag{4.4}$$

Eq. (4.4) implies that if $\beta_M < 0$, then for a given $E_b/N_0 > \gamma_0$ the *word* error probability of the concatenated code decreases to zero as the input block size is increased. This is summarized by saying that there is *word error probability interleaving gain.*[1]

In [7], we discuss the calculation of $\alpha(w,h)$ and β_M for a concatenated system of the type depicted in Figure 1, using analytical tools introduced in [3] and [4]. For example, for the parallel concatenation of q codes, with $q-1$ interleavers, we have

$$\beta_M \leq -q+2,$$

with equality if and only if each of the component codes is recursive. For a "classical" turbo code with $q = 2$, we have $\beta_M = 0$, so there is no word error probability interleaving gain. This suggests that the word error probability for classic turbo codes will not improve with input block size, which is in agreement with simulations.

---

[1] There is a similar conjecture for the bit error probability which we do not discuss in this paper. Suffice it to say that the interleaving gain exponent for bit error probability is $\beta_M - 1$.

As another example, consider the serial concatenation of two convolutional codes. If the inner code is recursive then,

$$\beta_M \le -\left\lfloor \frac{d^o_{\text{free}}+1}{2} \right\rfloor + 1,$$

where d^o_{free} is the minimum distance of the outer code. Therefore, for serial concatenated codes, if $d^o_f \ge 3$ there is interleaving gain for word error probability. (If the inner code is nonrecursive $\beta_M \ge 0$ and there is no interleaving gain.)

5. A Class of Simple Turbo-Like Codes.

In this section we will introduce a class of turbo-like codes which are simple enough so that we can prove the IGE conjecture. We call these codes *repeat and accumulate* (RA) codes. The general idea is shown in Figure 3. An information block of length N is repeated q times, scrambled by an interleaver of size qN, and then encoded by a rate 1 *accumulator*. The accumulator can be viewed as a truncated rate-1 recursive convolutional encoder with transfer function $1/(1+D)$, but we prefer to think of it as a block code whose input block $[x_1, \ldots, x_n]$ and output block $[y_1, \ldots, y_n]$ are related by the formula

$$\begin{aligned} y_1 &= x_1 \\ y_2 &= x_1 + x_2 \\ y_3 &= x_1 + x_2 + x_3 \\ &\vdots \\ y_n &= x_1 + x_2 + x_3 + \cdots + x_n. \end{aligned} \tag{5.1}$$

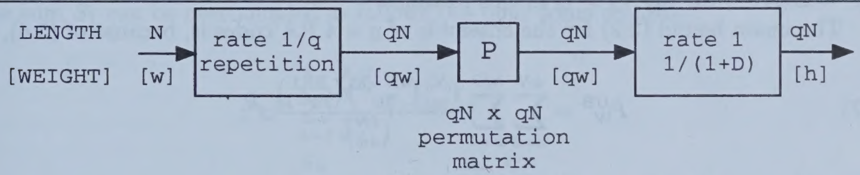


Figure 3. Encoder for a (qN, N) repeat and accumulate code. The numbers above the input-output lines indicate the length of the corresponding block, and those below the lines indicate the weight of the block.

To apply the union bound from Section 2 to the class of RA codes, we need the input-output weight enumerators for both the (qn, n) repetition code, and the (n, n) accumulator code. The outer repetition code is trivial: if the input block has length n, we have

$$A^{(o)}_{w,h} = \begin{cases} 0 & \text{if } h \ne qw \\ \binom{n}{w} & \text{if } h = qw. \end{cases} \tag{5.2}$$

The inner accumulator code is less trivial, but it is possible to show that (again assuming the input block has length n):

$$A^{(i)}_{w,h} = \binom{n-h}{\lfloor w/2 \rfloor}\binom{h-1}{\lceil w/2 \rceil - 1}. \tag{5.3}$$

It follows then from the general formula (3.1), that for the (qN, N) RA code represented by Figure 3, the ensemble IOWE is

$$\begin{aligned} A^{(N)}_{w,h} &= \sum_{h_1=0}^{qN} \frac{A^{(o)}_{w,h_1} A^{(i)}_{h_1,h}}{\binom{qN}{qw}} \\ &= \frac{\binom{N}{w}\binom{qN-h}{\lfloor qw/2 \rfloor}\binom{h-1}{\lceil qw/2 \rceil - 1}}{\binom{qN}{qw}}. \end{aligned} \tag{5.4}$$

From (5.4) it is easy to compute the parameters $\alpha(w,h)$ and β_M in (4.2) and (4.3). The result is

$$\alpha(w,h) = -\left\lceil \frac{(q-2)w}{2} \right\rceil \tag{5.5}$$

$$\beta_M = -\left\lceil \frac{(q-2)}{2} \right\rceil. \tag{5.6}$$

It follows from (5.6) that an RA code can have word error probability interleaving gain only if $q \geq 3$.

We are now prepared to use the union bound to prove the IGE conjecture for RA codes. In order to simplify the exposition as much as possible, we will assume for the rest of this section that $q = 4$, the extension to arbitrary $q \geq 3$ being straightforward but rather lengthy. For $q = 4$, (5.6) becomes $\beta_M = -1$, so the IGE conjecture is $P^{\mathrm{UB}}_W = O(N^{-1})$ for $E_b/N_0 > \gamma_0$ in this instance.

The union bound (2.2) for the ensemble of $q = 4$ RA codes is, because of (5.4),

$$P^{\mathrm{UB}}_W = \sum_{h=2}^{4N} \sum_{w=1}^{h/2} \frac{\binom{N}{w}\binom{4N-h}{2w}\binom{h-1}{2w-1}}{\binom{4N}{4w}} z^h. \tag{5.7}$$

Denote the (w,h)th term in the sum (5.7) by $T_N(w,h)$:

$$T_N(w,h) \overset{\mathrm{def}}{=} A_{w,h} z^h = \frac{\binom{N}{w}\binom{4N-h}{2w}\binom{h-1}{2w-1}}{\binom{4N}{4w}} z^h.$$

Using standard techniques (e.g. [8, Appendix A]), it is possible to show that for all (w,h),

$$T_N(w,h) \leq D 2^{h[F(x,y) + \log_2 z]}, \tag{5.8}$$

where $D = 4/\sqrt{\pi}$ is a constant, $x = w/4N$, $y = h/4N$,

$$F(x,y) = \frac{-\frac{3}{4}H_2(4x) + (1-y)H_2(\frac{2x}{1-y}) + yH_2(\frac{2x}{y})}{y},$$

and $H_2(x) = -x\log_2(x) - (1-x)\log_2(1-x)$ is the binary entropy function. The maximum of the function $F(x,y)$ in the range $0 \le 2x \le y \le 1-2x$ occurs at $(x,y) = (0.100, 0.371)$ and is 0.562281, so that if $\log_2 z < -0.562281$, the exponent in (5.8) will be negative.

Let us therefore assume that $\log_2 z < -0.562281$, which is equivalent to $E_b/N_0 = -(1/r)\ln z = -4\ln z \ge 4 \cdot \ln 2 \cdot 0.562281 = 1.559 = 1.928$ dB. If E is defined to be $E = -\log_2 z + 0.562281$, it follows from (5.8) for all w and h,

$$T_N(w,h) \le D2^{-hE}. \tag{5.9}$$

What (5.9) tells us is that if $E_b/N_0 > 1.928$ dB, most of the terms in the union bound (5.7) will tend to zero rapidly, as $N \to \infty$. The next step in the proof is to break the sum in (5.7) into two parts, corresponding to those terms for which (5.9) is helpful, and those for which it is not. To this end, define

$$h_N \overset{\text{def}}{=} \frac{3}{E}\log_2 N,$$

and write

$$\begin{aligned}
P_W^{\text{UB}} &= \sum_{h=2}^{4N}\sum_{w=1}^{h/2} T_N(w,h) \\
&= \sum_{h=2}^{h_N}\sum_{w=1}^{h/2} T_N(w,h) + \sum_{h=h_N+1}^{4N}\sum_{w=1}^{h/2} T_N(w,h) \\
&= S_1 + S_2.
\end{aligned}$$

It's easy to verify that when N is large enough, $A_{w+1,h}/A_{w,h} < 1$ for $h \le h_N$ and $w \le h/2 \le h_N/2$, which shows $A_{w,h}$ is a decreasing function of w for large N. Thus the sum S_1 can be overbounded as follows (we omit some details):

$$\begin{aligned}
S_1 &= \sum_{h=2}^{h_N}\sum_{w=1}^{h/2} T_N(w,h) \\
&= \sum_{h=2}^{h_N} T_N(1,h) + \sum_{h=2}^{h_N}\sum_{w=2}^{h/2} T_N(w,h) \\
&= O(N^{-1}) + \sum_{h=2}^{h_N}\sum_{w=2}^{h/2} T_N(w,h) \\
&\le O(N^{-1}) + \sum_{h=2}^{h_N}\sum_{w=2}^{h/2} A_{2,h}z^h \\
&= O(N^{-1}) + \sum_{h=2}^{h_N}\sum_{w=2}^{h/2} O(h^3/N^2)z^h \\
&= O(N^{-1}) + O(h_N^5/N^2) \\
&= O(N^{-1}).
\end{aligned}$$

For the sum S_2, we bound each term $T_N(w,h)$ by (5.9):

$$\begin{aligned}
S_2 &= \sum_{h=h_N+1}^{4N} \sum_{w=1}^{h/2} T_N(w,h) \\
&\le \sum_{h_N+1}^{4N} \sum_{w=1}^{h/2} D2^{-hE} \\
&= D/2 \sum_{h_N+1}^{4N} h2^{-hE} \\
&\le D\frac{2^{-Eh_N}(h_N+1)}{(1-2^{-E})^2} \\
&= O(N^{-3}\log_2 N) \\
&= o(N^{-2}).
\end{aligned}$$

We have therefore shown that for the ensemble of $q = 4$ RA codes, if $E_b/N_0 >$ 1.928 dB,

$$P_W^{\mathrm{UB}} = S_1 + S_2 = O(N^{-1}) + o(N^{-1}) = O(N^{-1}), \tag{5.10}$$

which as we saw above, is the IGE conjecture in this case.

Although the union bound gives a proof of the IGE conjecture for RA codes, the resulting value of γ_0 is by no means the best possible. Indeed, if we use the recent Viterbi-Viterbi improved union bound [9] to bound the sum S_2, we can lower the value of γ_0 considerably, e.g. for $q = 4$ from 1.928 dB to 0.313 dB. In Figure 4 and Table 1 we display our numerical results on RA codes. There we compare the "cutoff threshold" γ_0 for RA codes with q in the range $3 \le q \le 8$ using both the classical union bound and the Viterbi-Viterbi improved union bound to the cutoff threshold for the ensemble of all codes (i.e., "random codes") of a fixed rate. We believe that these values of γ_0 can be reduced still further, for example by using the bound of [6] instead of the Viterbi-Viterbi bound.

q	3	4	5	6	7	8
RA Codes (Union Bound)	2.200	1.928	1.798	1.721	1.670	1.631
Random Codes (Union Bound)	2.031	1.853	1.775	1.694	1.651	1.620
RA Codes (Viterbi Bound)	1.112	0.313	−0.125	−0.402	−0.592	−0.731
Random Codes (Viterbi Bound)	0.214	−0.224	−0.486	−0.662	−0.789	−0.885
Binary Shannon Limit	−0.495	−0.794	−0.963	−1.071	−1.150	−1.210

Table 1. Numerical data gleaned from Figure 4.

6. Performance of RA Codes with Iterative Decoding.

The results of this paper show that the performance of RA codes *with maximum-likelihood decoding* is very good. However, the complexity of ML decoding of RA

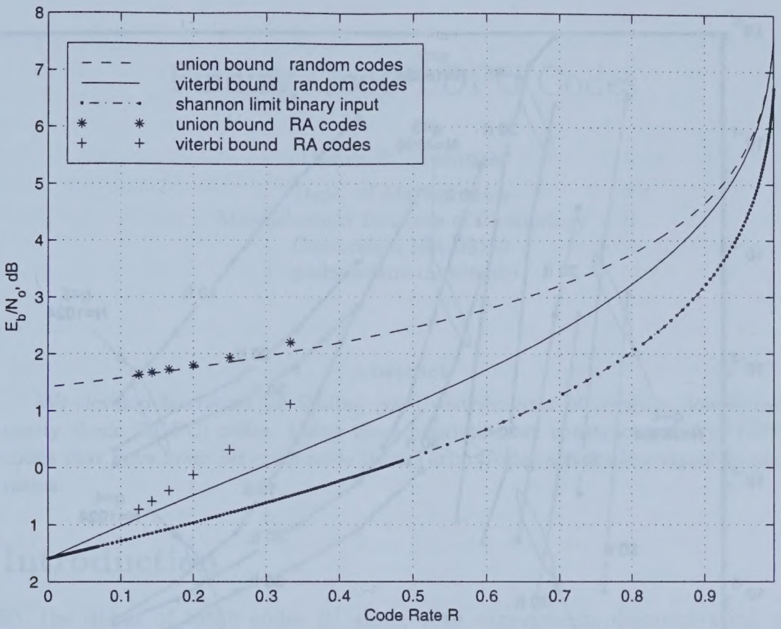


Figure 4. Comparing the RA code "cutoff threshold" to the cutoff rate of random codes using both the classical union bound and the Viterbi-Viterbi improved union bound.

codes, like that of all turbo-like codes, is prohibitively large. But an important feature of turbo-like codes is the availability of a simple iterative, message passing decoding algorithm that approximates ML decoding. We wrote a computer program to implement this "turbo-like" decoding for RA codes with $q = 3$ (rate 1/3) and $q = 4$ (rate 1/4), and the results are shown in Figure 5. We see in Figure 4, for example, that the empirical cutoff threshold for RA codes for $q = 3$ appears to be less than 1 dB, compared to the upper bound of 1.112 dB found in Table 1.

References.

1. C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: turbo codes," *Proc. 1993 IEEE International Conference on Communications*, Geneva, Switzerland (May 1993), pp. 1064–1070.
2. S. Benedetto and G. Montorsi, "Unveiling turbo codes: some results on parallel concatenated coding schemes", *IEEE Trans. on Inf. Theory*, vol. 42, no. 2 (March 1996), pp. 409–428..
3. S. Benedetto and G. Montorsi, "Design of parallel concatenated convolutional codes," *IEEE Transactions on Communications*, vol. 44, no. 5, (May 1996) pp. 591–600.
4. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "Serial concatenation of interleaved codes: performance analysis, design, and iterative decoding," *IEEE Trans. on Information Theory*, vol. 44, no. 3, (May 1998), pp. 909–926.

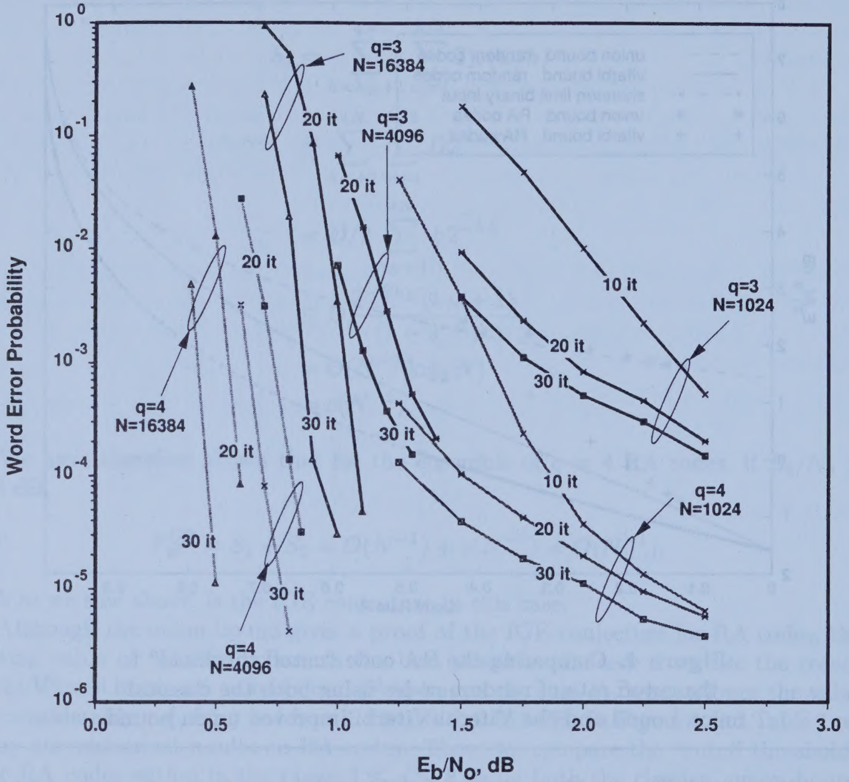


Figure 5. Simulated performance of iterative decoding of RA codes on an AWGN channel.

5. D. Divsalar and F. Pollara, "On the design of turbo codes," *TDA Progress Report* vol. 42-123 (November 15, 1995), pp. 99–121.
6. S. Dolinar, L. Ekroot, and F. Pollara, "Improved Error Probability Bounds for Block Codes for the Gaussian channel," Proc. 1994 ISIT, p. 243.
7. D. Divsalar and R. J. McEliece, "On the design of concatenated coding systems with interleavers," JPL TMO Progress Report vol. 2-134 (August 15, 1998), pp. 1–22. (`http://tmo.jpl.nasa.gov/tmo/progress_report/42-134/134D.pdf` .)
8. W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes, 2nd. ed.* Cambridge, Mass.: The MIT Press, 1972.
9. A. J. Viterbi and A. M. Viterbi, "Improved union bound on linear codes for the input-binary AWGN channel, with applications to turbo decoding," Proc. Winter 1998 Information Theory Workshop, San Diego California, Feb. 1998, p. 72.

Finding Good LDPC Codes

Daniel A. Spielman*
Dept. of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139
spielman@math.mit.edu

Abstract

We develop heuristics for finding good distributions of irregular low density parity check (LDPC) codes. Using these heuristics, we construct rate 1/2 LDPC codes that have lower bit error rates than Turbo Codes at very low signal-to-noise ratios.

1 Introduction

In 1993, the debut of turbo codes [1] along with experiments demonstrating their remarkably low bit-error-rates at very low signal-to-noise ratios excited the coding theory community. A few years later, low density parity check codes [2] reemerged through experiments demonstrating [7, 8] that they too allow low bit-error-rates at low signal-to-noise ratios, although not as low at turbo codes. Since then, LDPC codes have been catching up. Improvements over Gallager's initial construction were made by using carefully chosen irregular graphs [6] and by using fields other than GF(2) [9]. Recently, Davey and MacKay [10] have constructed irregular rate 1/4 LDPC codes over GF(8) that beat Turbo Codes at very low SNRs. In this work, we construct rate 1/2 irregular LDPC codes that beat Turbo Codes at very low SNRs. Our codes are designed to work over $GF(2)$, so we expect to see better performance after we incorporate larger fields into our techniques.

Our codes were constructed by heuristic algorithms designed to find good degree sequences for irregular LDPC codes. These algorithms are a generalization of those used in [5]. As our present implementations of these algorithms are quite slow, we have not yet had enough time to understand their full potential. We expect such understanding to occur only after we have optimized the implementations of the heuristics. The codes presented in this paper were the result of applying one heuristic once, and another twice. This took several days.

All results in this paper are presented for the Gaussian channel.

1.1 Irregular Codes

A LDPC code is specified by a bipartite graph between a set of n *symbol nodes* and a set of $(1-r)n$ *check nodes*, where n is the block length of the code and r is the designed

---

*Supported in part by NSF CAREER Award CCR-9701304, an Alfred P. Sloan Foundation Fellowship, and a Sloan/Cabot Award from the M.I.T. School of Science

rate[1] Each symbol node is associated with a bit to be transmitted, and the check nodes impose linear constraints upon those bits.

When Gallager[2] introduced LDPC codes, he proposed using bipartite graphs in which the degrees of nodes on each side of the graph were uniform (*e.g.*, all the symbol nodes could could have degree 3 and all the check nodes could have degree 6). We will call such LDPC codes "regular".

For this work, we will build LDPC codes from the types of irregular graphs suggested in [4, 5]. These graphs are specified by a symbol node degree sequence $(\lambda_1, \lambda_2, \lambda_3, \ldots)$ and a check node degree sequence $(\rho_1, \rho_2, \rho_3, \ldots)$, where λ_i (ρ_i) specifies the proportion of symbol nodes (check nodes) of degree i, and

$$\sum_i \lambda_i = 1, \qquad \sum_i \rho_i = 1.$$

Given two degree sequences, we form a LDPC code by choosing a random bipartite graph in which the proportions of nodes of each degree is given by the appropriate term in one of the degree sequences.

In [4], we proved that, by carefully choosing the degrees of nodes on each side of the graph, it was possible to produce LDPC codes that quickly approached the capacity of the erasure channel. In [6], we presented experiments which showed that some irregular LDPC codes from [4] performed better under belief propagation decoding than any regular LDPC codes. In [5], we analyzed the performance of a discretized version of belief propagation, and showed that irregular LDPC codes performed better than regular LDPC codes when decoded using this algorithm. The main tool of [5] was a linear programing based algorithm that, given a degree sequence for the check nodes, would find an optimal degree sequence for the symbol nodes, and *vice versa*. In this work, we adapt this approach to full belief propagation. However, to make this adaptation, we must sacrifice guarantees of optimality.

1.2 Belief Propagation

In this section, we sketch how the belief propagation algorithm works. In this description, we assume that all codewords are equally likely and we only associate one bit with each symbol node. We refer readers desiring a more complete description to [7, 12].

The belief propagation algorithm (a.k.a. the sum-product algorithm) begins by considering the received symbols and computing, for each symbol node, the probability that the associated bit is 0 given the received symbol for that bit. Each edge in the graph is then assigned the value of the symbol node to which it is attached. After this initialization, the algorithm proceeds by alternating between *check passes* and *message passes*.

We understand a check pass by imagining a random 0/1 variable associated with each edge, with the value of the edge being the probability its variable is zero. During a check pass, each edge is assigned the value of the probability that the sum of the variables associated with the other edges connected to its check has parity zero, assuming that the variables are independent. In particular, the value an edge is assigned during a check pass is not a function of the value of that edge before the check pass. To make this explicit, consider a check node with values $p_1, \ldots, p_k$ on its edges. After the check pass, these edges will carry the values

---

[1] It is possible for the rate of the code to be higher than $(1-r)n$ if there are redundancies among the constraints imposed by the check nodes.

$$\oplus_{k-1}(p_2, \ldots p_k), \oplus_{k-1}(p_1, p_3, \ldots p_k), \ldots, \oplus_{k-1}(p_1, \ldots, p_{k-1}),$$

respectively, where

$$\oplus_{k-1}(p_1, \ldots, p_{k-1}) = \frac{1 + \prod_i (2p_i - 1)}{2}.$$

To understand a message pass, we imagine that each edge carries a random variable obtained by sending the bit associated with its symbol node through some channel, with the value of the edge being the probability that the bit was 0 zero given that the variable was received. During a message pass, each edge is assigned the probability that the bit associated with its symbol node is 0, assuming that the received symbol for that bit and the values carried by the other edges of that symbol node represent independent observations of the bit. For a symbol node, if p_0 is the probability that the bit associated with the node is zero given the received symbol for that bit, and if the values on the edges connected to that symbol node are $p_1, \ldots p_k$, then the values on those edges after the message pass will be

$$\otimes_k(p_0, p_2, \ldots, p_k), \otimes_k(p_0, p_1, p_3, \ldots, p_k), \ldots \otimes_k (p_0, p_1, \ldots, p_{k-1}),$$

where

$$\otimes_k = \frac{\prod_i p_i}{\prod_i p_i + \prod_i (1 - p_i)}.$$

It will later be useful to note that

$$\otimes_k(p_0, p_1, \ldots p_k) = \otimes_2(p_0, \oplus_{k-1}(p_1, \ldots p_{k-1})).$$

After some number of rounds, the algorithm is stopped after a check pass. The output of the algorithm is obtained by assigning each symbol node the best estimate of its bit's value that can be obtained from the symbol received for that bit and the values of the edges connected to that symbol node, assuming that each of these values comes from an independent measurement. In the language of the previous paragraph, this quantity is $\otimes_{k+1}(p_0, p_1, \ldots p_k)$.

Unless the graph underlying the LDPC code is a tree, the above independence assumptions can be false. When the graph is a tree, this algorithm is known to converge to the best a posteriori estimate for each bit[11, 12]. In experiments with randomly generated LDPC codes, it has been observed that this algorithm converges to a very good estimate for each bit, even though the underlying graph is not a tree. This observation has led to the conjecture that the performance of the belief propagation algorithm on LDPC codes derived from random graphs is almost identical to the performance of the belief propagation algorithm on an infinite tree, and that it becomes more similar as the block size of the LDPC code grows. One justification for this conjecture is that the expected girth of a random graph with fixed degree sequences grows logarithmically in the number of nodes in the graph. Thus, the first few rounds of belief propagation decoding do accurately reflect the performance of belief propagation decoding in an infinite tree.

While it is known that the belief propagation algorithm does converge on infinite trees, we are unaware of any simple formula for what it converges to. That is, given the structure of the infinite tree, given an initial SNR, and given a number of rounds, we know of no simple way of determining the bit error rate after that number of rounds. In fact, we are unable to do this for **any** non-trivial[2] pair of degree sequences even as n goes to infinity!

However, given a pair of degree sequences, an initial SNR and a number of rounds, we can sample the distribution of values on edges after that number of rounds. For convenience, we assume that the all-0 codeword was transmitted. Let M_0 denote the distribution of values on edges after the initialization of the belief propagation algorithm. Let C_1 denote the distribution after the first check pass, M_1 denote the distribution after the first message pass, and so on. We can sample M_0 by simulating the Gaussian channel. We can sample C_i by performing check passes on samples from M_{i-1}, and we can sample M_i by performing message passes on samples from C_{i-1}, and so on.

To make this precise, it helps to first consider the infinite tree corresponding to a regular graph. Assume that every symbol node has degree l and that every check node has degree r. Then, we sample from C_i by drawing $r-1$ samples from M_{i-1}, say $p_1, \ldots p_{r-1}$, and computing

$$\otimes_{r-1}(p_1, \ldots p_{r-1}).$$

Similarly, we sample from M_i by drawing $l-1$ samples from C_{l-1}, say $p_1, \ldots p_{l-1}$ and one sample from M_0, say p_0, and computing

$$\oplus_l(p_0, p_1, \ldots p_{l-1}).$$

We will overload our notation by expressing these relations between distributions as

$$\begin{aligned} C_i &= \bigoplus_{r-1}(\overbrace{M_{i-1}, \ldots, M_{i-1}}^{r-1}), \textit{and} \\ M_i &= \bigotimes_{l}(M_0, \overbrace{C_i, \ldots C_i}^{l-1}). \end{aligned}$$

If we apply this approach to sampling precisely, then we are actually simulating the process on a very large tree. As this requires immense amounts of time, we instead approximate these distributions by generating a large set of samples from each. Whenever we need a sample from a distribution, we just pick one at random from the pre-sampled set. Naturally, we form our set of samples corresponding to C_i before those corresponding to M_i, and so on. For convenience, we will assume that our set of samples for each space approximates that which we would get if we chose independent samples from that space.

As the correct definition of the infinite tree corresponding to an irregular graph specified by a pair of degree sequences is not transparent to the author, we will settle for an explanation of how to sample from the corresponding distributions C_i and M_i. First, note that the probability that a randomly chosen edge is connected to a symbol node of degree i is

$$\frac{(\lambda_i \cdot i)}{\sum_j (\lambda_j \cdot j)}.$$

It is important to note that this probability is very different from the probability that a randomly chosen symbol node has degree i. For convenience, set $\lambda'_i = (i\lambda_i)/(\sum_j j\lambda_j)$ and $\rho'_i = (i\rho_i)/(\sum_j j\rho_j)$. To sample from M_i we, with probability λ'_j, compute $\otimes_j(p_0, p_1, \ldots, p_{j-1})$, where p_0 is drawn from M_0 and $p_1, \ldots p_{j-1}$ are drawn from C_j. We write this relation as:

---

[2] We would consider the graphs in which symbol nodes have degree 3 and check nodes have degree 6 to be non-trivial.

$$\begin{aligned} M_i &= \sum_j \lambda'_j \bigotimes_j (M_0, \overbrace{C_j, \ldots, C_j}^{j-1}), \text{ and} && (1) \\ C_i &= \sum_j \rho'_j \bigoplus_{j-1} (\overbrace{M_{j-1}, \ldots, M_{j-1}}^{j-1}), && (2) \end{aligned}$$

where by a linear combination of probability spaces with coefficient sum one, we mean that each is sampled with probability according to its coefficient.

2 The LP Technique

The goal of the linear programming technique[5] is, given one degree sequence and an initial noise level, to find a complementary degree sequence for which the probability of bit error goes to zero as the number of decoding rounds increases. In[5] this task was simplified by the fact that the distributions M_i and C_i produced by the discretized belief propagation algorithm could be completely specified by one parameter. Moreover, the parameter specifying the distribution M_i (C_i) could be obtained from a simple formula in the parameter specifying the distribution C_i (M_{i-1}).

In our case, we cannot simply characterize the distributions M_i and C_i encountered in the belief propagation algorithm. Instead, we work with sets of samples believed to approximate these distributions as described in the previous section.

Given a check node degree sequence and an initial signal-to-noise ratio, we want to set up a linear program that will find a symbol node degree sequence such that the distribution M_i is better than the distribution M_{i-1} for all i, if such a distribution exists. To measure the quality of a distribution, we use the objective function:

$$F(M_i) = -\int_p log(p) d\mu_i(p), \qquad (3)$$

where μ_i is the measure associated with distribution M_i. The smaller $F(M_i)$ is, the better we consider M_i to be. We work with (3) because it is easy to approximate from our set of samples and because it is clearly respects formula (1). That is,

$$-\int_p log(p) d\mu_i(p) = -\sum_j \lambda'_j \int_p log(p) d\mu_{i,j}(p),$$

where $\mu_{i,j}$ is the measure corresponding to the distribution

$$M_{i,j} = \bigotimes_j (M_0, \overbrace{C_i, \ldots, C_i}^{j-1}).$$

It is almost possible to construct a linear program that, given M_0 and a check node degree sequence, will solve for λ_is so that $F(M_i) < F(M_{i-1})$ for all i. Unfortunately, the dependence of M_i upon M_{i-1} through C_i prevents us from solving these inequalities simultaneously.

3 Our Heuristics

We tried two approaches to finding good degree sequences. These were:

(a) **ignore the dependencies:** In this approach, we fix symbol and check degree sequences, generate the corresponding distributions $M_0, \ldots M_{k-1}$ and $C_1, \ldots C_k$ for some k, and then try to find a symbol (check) degree sequence that would do better than the original one with input distributions $C_1, \ldots, C_k$ $(M_0, \ldots M_{k-1})$. That is, given $(\lambda'_i)_i$ and $(\rho'_i)_i$, we try to find $(\alpha'_i)_i$ so that

$$\sum_i \lambda'_i F(\bigotimes_i(M_0, \overbrace{C_j, \ldots, C_j}^{i-1})) > \sum_i \alpha'_i F(\bigotimes_{i-1}(M_0, \overbrace{C_j, \ldots, C_j}^{i-1})),$$

for all j, while keeping the rate of the resulting code the same.

(b) **Pretend that all distributions are like the input distribution:** For any initial signal-to-noise ratio, we get an initial distribution M_0. Construct many such initial distributions $M_{0,l}$ corresponding to increasing signal-to-noise ratios as l grows. For some check degree sequence $(\rho'_i)_i$, use linear programming to find a degree sequence $(\lambda'_i)_i$ such that

$$\sum_i \lambda'_i F(\bigotimes_i(M_0, \overbrace{C_l, \ldots C_l}^{i-1})) < F(\bigotimes_2(M_0, M_{0,l})),$$

where

$$C_l = \sum_i \rho'_i \bigoplus_{i-1}(\overbrace{M_{0,l}, \ldots, M_{0,l}}^{i-1}).$$

That is, we just try to find symbol degree sequences such that if the distributions after message passes looked like the initial distributions, each successive round would have better distributions.

4 Results

The experiments reported here were done on LDPC codes built from degree sequences obtained by combining approaches (a) and (b). We first used approach (b) to find a good symbol degree sequence given the check degree sequence $\rho_9 = 1$. We then used approach (a) to refine this symbol degree sequence, keeping the same check degree sequence. We then refined the symbol degree sequence once more. The symbol node degree sequence we settled on was $(\lambda_2 = .4804, \lambda_3 = .2797, \lambda_6 = .0138, \lambda_7 = .1830, \lambda_{31} = .0431)$.

in Figure 1 and Figure 2, we show the data obtained in simulations of rate 1/2 codes of lengths 15824 and 32242 respectively. In Figure 4, we compare the performance of these two codes with that of Turbo codes reported in [3]. The points on the lowest curve at .6 and .7 db should be viewed with suspicion until we have the results of more simulations. The same holds for the highest point at 8 db. We expect that when we perform more simulations with more decoding rounds in each trial, the middle columns of Figures 1 and 2 will better resemble the last columns of these Figures.

SRN Db	rounds	trials	BER	< 15700	$< 13K$	$< 11K$	$< 10K$
.4	250	500	.1627	298	243	146	55
.45	250	1500	.0918	556	402	227	83
.5	150	3000	.0505	694	429	233	89
.6	150	2000	.0098	111	55	29	11
.7	120	6000	.00104	52	12	7	3
.8	120	6000	1.29e-04	8	0	0	0

Figure 1: This code has block length 15842. For each signal-to-noise level, we list the number of rounds of belief propagation that were used to produce these results, the number of trials at this noise level, the bit error rate over all these trials, and the number of times the decoding algorithm got less that 15700, 13000, 11000, and 10000 of the bits correct.

SRN Db	rounds	trials	BER	$< 32K$	$< 27K$	$< 24K$	$< 20K$
.4	250	500	.0831	183	134	81	8
.45	250	1500	.0371	265	175	103	6
.5	150	1200	.00849	68	28	16	3
.6	150	2000	3.38e-04	3	1	1	0
.7	120	7400	1.04e-04	6	0	0	0
.7	150	4000	1.94e-04	3	1	1	1

Figure 2: This code has block length 32242. For each signal-to-noise level, we list the number of rounds of belief propagation that were used to produce these results, the number of trials at this noise level, the bit error rate over all these trials, and the number of times the decoding algorithm got less that 32000, 27000, 24000, and 20000 of the bits correct.

5 Failed Attempts

We also tried to use approach (a) to obtain better check degree sequences. While the linear program did return more interesting check degree sequences, these did not appear to perform as well in our simulations. In fact, we observed problematic behavior: some of the codes we produced would make quick progress initially, but then slow down to a crawl and occasionally get stuck with very few errors remaining.

However, we should note that we only made a few attempts at finding good check degree sequences, and it might be that if we tried a few more times we would have better luck.

6 Conclusions and Future Work

At the moment, we are unsure whether this approach can be used to find substantially better LDPC codes over $GF(2)$. So far, the work has progressed slowly

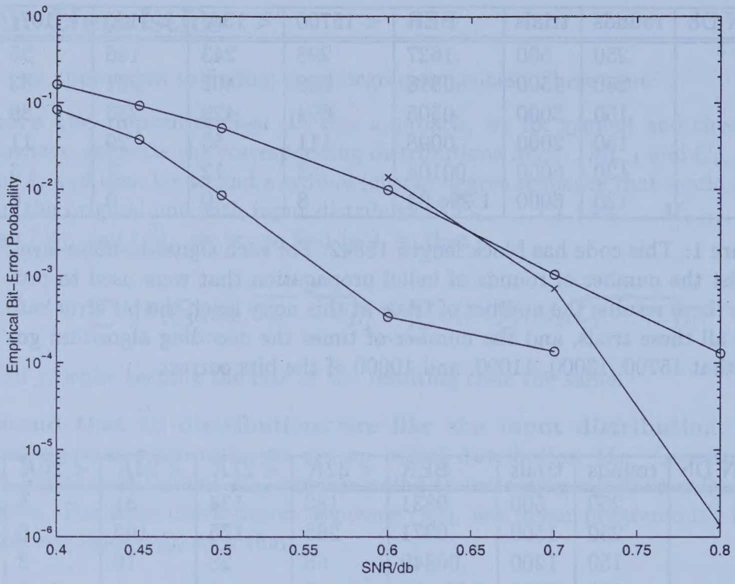


Figure 3: The leftmost two curves were generated from the data in Figures 1 and 2. The other curve is from the 1/2 rate Turbo code of [3] The points on the lowest curve at .6 and .7 db should be viewed with suspicion until we have the results of more simulations. The same holds for the highest point at 8 db.

because we implemented our algorithms in Matlab. This was quite slow: just obtaining data for approach (a) or (b) with reasonable numerical accuracy could take a whole day. We are presently translating the algorithms to C, and are observing a dramatic speedup. This should make it possible to explore these approaches more fully.

The obvious next step in this research is to apply these techniques to the construction of codes over fields other than $GF(2)$. We will do this as soon as we have completed our faster implementation. Given how much better regular LDPC codes over $GF(8)$ are than regular codes over $GF(2)$, and given how much better these irregular LDPC codes over $GF(2)$ are than the regular codes over $GF(2)$, we presume that terrific results will result from combining the two appraoches.

References

[1] C. Berrou, A Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes", *Proceedings of IEEE International Communications Conference, 1993.*

[2] R. G. Gallager, **Low-Density Parity-Check Codes**, MIT Press, 1963.

[3] JPL. Turbo code performance. Available from **http://www331.jpl.nasa.gov/public/TurboPerf.html**, September 1998.

[4] M. Luby, M. Mitzenmacher, M. A. Shokrollahi, D. A. Spielman, and V. Stemann, "Practical Loss-Resilient Codes", *Proc.* 29^{th} *ACM Symposium on Theory of Computing*, 1997, pp. 150–159.

[5] M. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, "Analysis of Low Density Codes and Improved Designs Using Irregular Graphs", *Proceedings of the* 30^{th} *ACM Symposium on Theory of Computing*, 1998, pp. 249–258.

[6] M. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, "Improved Low Density Parity Check Codes Using Irregular Graphs and Belief Propagation", submitted to the *1998 International Symposium on Information Theory.*

[7] D. J. C. MacKay and R. M. Neal, "Good Error Correcting Codes Based on Very Sparse Matrices", to appear in *IEEE Transactions on Information Theory.* Available from **http://wol.ra.phy.cam.ac.uk/mackay**.

[8] D. J. C. MacKay and R. M. Neal, "Near Shannon Limit Performance of Low Density Parity Check Codes", to appear in *Electronic Letters.*

[9] M. C. Davey and D. J. C. MacKay, "Low Density Parity Check Codes over $GF(q)$", early manuscript, 1997.

[10] M. C. Davey and D. J. C. MacKay, "Low Density Parity Check Codes over $GF(q)$", *Information Theory Workshop 1998, Killarney, Ireland.*

[11] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, 1988.

[12] N. Wiberg, "Codes and decoding on general graphs" Ph.D. dissertation, Dept. Elec. Eng, U. Linköping, Sweeden, April 1996.

Comparison of Constructions of Irregular Gallager Codes

David J. C. MacKay, Simon T. Wilson and Matthew C. Davey
Department of Physics, University of Cambridge
Cavendish Laboratory, Madingley Road,
Cambridge, CB3 0HE, United Kingdom.
mackay|stw11|mcdavey@mrao.cam.ac.uk

In Proceedings of the 1998 Allerton Conference on Communication, Control, and Computing.
Submitted to IEEE Transactions on Communications 30 July 1998.

Abstract

The low density parity check codes whose performance is closest to the Shannon limit are 'Gallager codes' based on irregular graphs. We compare alternative methods for constructing these graphs and present two results.

First, we find a 'super-Poisson' construction which gives a small improvement in empirical performance over a random construction.

Second, whereas Gallager codes normally take N^2 time to encode, we investigate constructions of regular and irregular Gallager codes which allow more rapid encoding and have smaller memory requirements in the encoder. We find that these 'fast-encoding' Gallager codes have equally good performance.

1 Introduction

Gallager codes [3, 4] are low density parity check codes constructed at random subject to constraints on the weight of each row and of each column. The original *regular* Gallager codes have very sparse random parity check matrices with uniform weight t per column and t_r per row. [We will also use the term 'regular' for codes which have nearly uniform weight columns and rows — for example, codes which have some weight 2 columns and some weight 3 columns.] These codes are asymptotically good, and can be practically decoded with Gallager's sum–product algorithm giving near Shannon limit performance when large block lengths are used [8, 9, 6]. Regular Gallager codes have also been found to be competitive codes for short block-length CDMA applications [10].

Recent advances in the performance of Gallager codes are summarised in figure 1. The rightmost curve shows the performance of a regular binary Gallager code with rate 1/4. The best known binary Gallager codes are *irregular* codes whose parity check matrices have *nonuniform* weight per column [5]; the performance of one such code is shown by the second curve from the right. The best known Gallager codes of all are Gallager codes defined over finite fields $GF(q)$ [2, 1]. The remaining two solid curves in figure 1 show the performance of a regular Gallager code over $GF(16)$ [2] and an irregular code over $GF(8)$ with bit error probability of 10^{-4} at $E_b/N_0 = -0.05$dB [1]. In comparing this

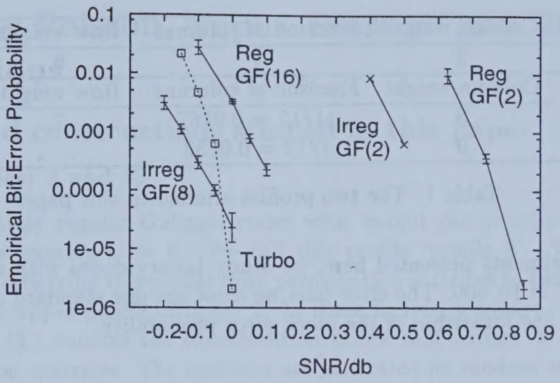


Figure 1: Empirical results for Gaussian Channel, Rate 1/4 Left–Right : Irregular LDPC, $GF(8)$ blocklength 24000 bits; JPL Turbo, blocklength 65536 bits; Regular LDPC, $GF(16)$, blocklength 24448 bits; Irregular LDPC , $GF(2)$, blocklength 64000 bits; Regular LDPC, $GF(2)$, blocklength 40000 bits. (Reproduced from [1].)

code with the rate 1/4 Turbo code shown by the dotted line, the following points should be noted. (1) The transmitted blocklength of the irregular Gallager code is only 24000 bits, whereas that of the Turbo code is 65536 bits. (2) The errors made by the Gallager code are (empirically) all detected errors, whereas Turbo codes make undetected errors.

The excellent performance of irregular Gallager codes is the motivation for this paper, in which we explore ways of further enhancing these codes.

The irregular codes of Luby, Mitzenmacher, Shokrollahi and Spielman [5] have parity check matrices with both nonuniform weight per row and nonuniform weight per column. It has not yet been established whether both of these nonuniformities are desirable. Our intuition is that, while it is beneficial to have nonuniform column weights, the weight per row should in fact be as uniform as possible, and in this paper we concentrate on irregular codes having this property.

We can define an irregular Gallager code in two steps. First, we select a *profile* that describes the desired number of columns of each weight and the desired number of rows of each weight. The parity check matrix of a code can be viewed as defining a bipartite graph with 'bit' vertices corresponding to the columns and 'check' vertices corresponding to the rows. Each non–zero entry in the matrix corresponds to an edge connecting a bit to a check. The profile specifies the degrees of the vertices in this graph.

Second, we choose a *construction method*, that is, a pseudo–random algorithm for putting edges between the vertices in a way that satisfies the constraints. [In the case of nonbinary Gallager codes we also need to choose an algorithm for assigning values to the non–zero entries in the matrix.]

This paper has two parts. In the first part (section 3) we compare alternative construction methods for a fixed profile in order to find out whether the construction method matters. In the second part (section 4) we examine regular and irregular constructions which lend themselves to rapid encoding. One motivation for this second study is that the only drawback of regular Gallager codes compared to Turbo codes for CDMA applications appears to be their greater encoding complexity [10].

	Column weight	Fraction of columns	Row weight	Fraction
Profile 3	3	1	6	1
Profile 93	Column weight	Fraction of columns	Row weight	Fraction
	3	$11/12 = 0.9167$	7	1
	9	$1/12 = 0.0833$		

Table 1: The two profiles studied in this paper

In the experiments presented here, we study binary codes with rate 1/2 and blocklength about $N = 10,000$. The error bars we show are one standard deviation error bars on the value of the logarithm of the block error probability.

2 Constructions

We mention the following methods.

Poisson: The edges are placed 'completely at random' subject to the profile constraints, and the rule that you can't put two edges between one pair of vertices, which would correspond to a double entry in the parity check matrix. One way to implement a Poisson construction is to make a list of all the columns in the matrix, with each column appearing in the list a number of times equal to its weight, then make a similar list of all the rows in the matrix, each row appearing with multiplicity equal to its weight, and then map one list onto the other by a random permutation, taking care not to create duplicate entries [5].

Poisson A: The edges are placed 'at random' as in Poisson, but with the added constraint that no two columns in the parity check matrix are allowed to have an overlap greater than one, *i.e.*, there are no cycles of length 4 in the graph. [Similar to construction 1A in [9].]

Poisson B: The edges are placed 'at random' as in Poisson A, but with the added constraint that the bipartite graph defined by the matrix should have no cycles of length less than some l. [This can be quite hard to enforce if the profile includes high weight rows or columns.] [Similar to construction 1B in [9].]

Cayley graphs. For some profiles and blocklengths, we can create bipartite graphs from Cayley graphs and use these to define a parity check matrix. These constructions are deterministic [7].

Permutations. We can build parity check matrices by superposing random permutation matrices. The convenience of this method depends on the profile. There are many ways of laying out these permutation matrices to satisfy a given profile. We will distinguish *'super-Poisson'* and *'sub-Poisson'* constructions.

- In a super-Poisson construction, the expected connectivity is more bunched than that of Poisson;
- in a sub-Poisson construction the expected connectivity is more uniform than Poisson.

3 Comparing Poisson, ‘super–Poisson’ and ‘sub–Poisson’ constructions

3.1 Profiles and constructions studied in this paper

3.1.1 Regular codes: `3` and `33`

As our baseline we study regular Gallager codes with weight per column exactly $t = 3$ and weight per row exactly $t_r = 6$. We call this profile ‘profile 3’. We construct parity check matrices satisfying this profile from permutation matrices in two ways shown diagramatically in the upper panels of figure 2. In these figures, a square containing an integer (for example, ‘3’) denotes the superposition inside that square of that number of random permutation matrices. The matrices are generated at random subject to the constraint that no two non–zero entries coincide.

3.1.2 Irregular codes: `93p`, `93a`, `93x` and `93y`

We chose the profile ‘93’ shown in table 1. It has columns of weight 9 and of weight 3; all rows have weight 7. Note that this profile only differs from the regular profile ‘3’ in that some extra 1s are added to $\frac{1}{12}$ of the columns. We emphasize that this profile has not been carefully optimized, so the results of this paper should not be taken as describing the best that can be done with irregular binary Gallager codes. We chose this profile because it lends itself to interesting experiments.

We will refer to the bits that connect to 9 checks as ‘elite’ bits. We use four different constructions that match this profile, named as follows. These constructions are depicted diagrammatically in the upper panels of figure 2.

Poisson: `93p`. In this construction, while most checks will connect to one or two elite bits, a fraction of them will connect to more than two elite bits, and some will connect to none.

Sub–Poisson: `93a`. This construction allocates exactly one or two elite bits to each check.

Super–Poisson: `93x` and `93y` are respectively moderately and very super–Poisson. In `93y`, one third of the checks are connected to *four* elite bits, one third are connected to *one*, and one third are connected to *none*.

3.2 Results

3.2.1 Variability within each construction

For each construction we created several codes in order to assess the variability of performance within each ensemble. The results are shown in figure 2. We see no significant variability among the `3`, `33`, `93i` (Poisson) or `93a` (sub–Poisson) codes. But among the super–Poisson codes, `93x` and `93y`, there is some variability, with some codes showing an error floor.

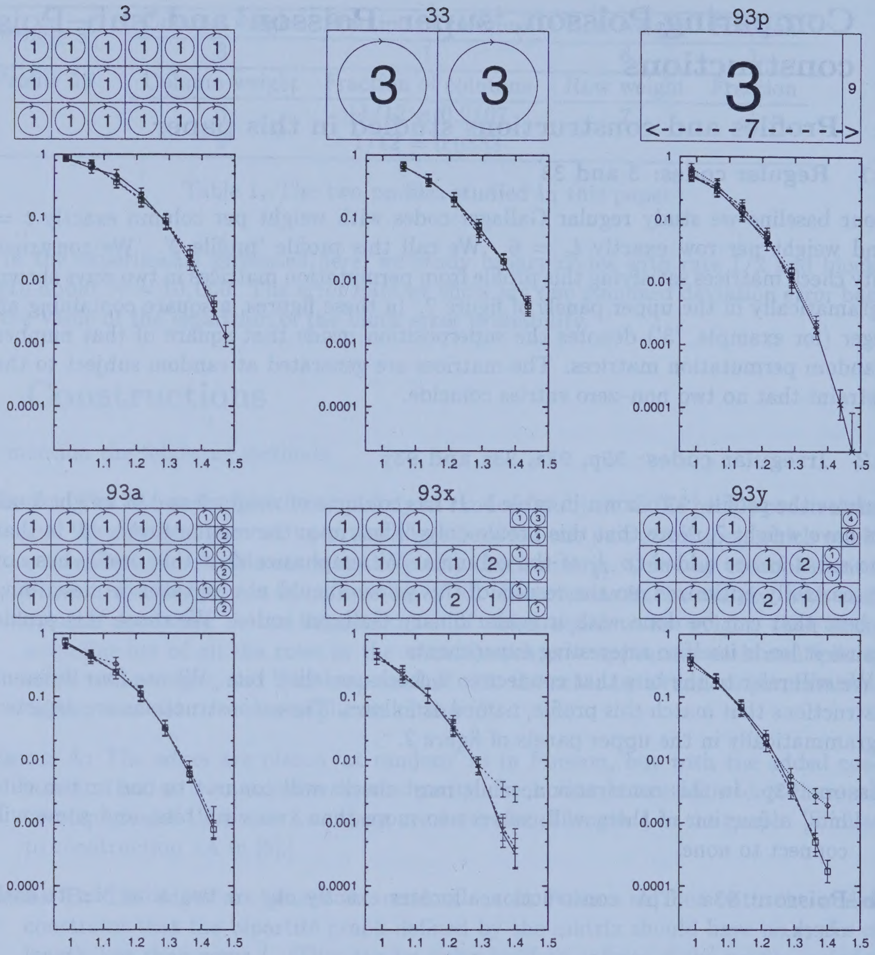


Figure 2: Upper panels: constructions of regular and irregular codes. Lower panels: performance of these codes. The construction types shown are regular, (3, 33), Poisson (93p), sub–Poisson (93a), super–Poisson (93x), and super–Poisson (93y).
Notation for upper panels for all constructions except 93p: an integer represents a number of permutation matrices superposed on the surrounding square. Horizontal and vertical lines indicate the boundaries of the permutation blocks. Notation for the Poisson construction 93p: integers '3' and '9' represent column weights. The integer '7' represents the row weight.
Lower panels show the performance of several random codes of each construction. Vertical axis: block error probability. Horizontal axis: E_b/N_0 in dB. All codes have $N = 9972$, and $K = M = 4986$.
All errors were detected errors, as is usual with Gallager codes.

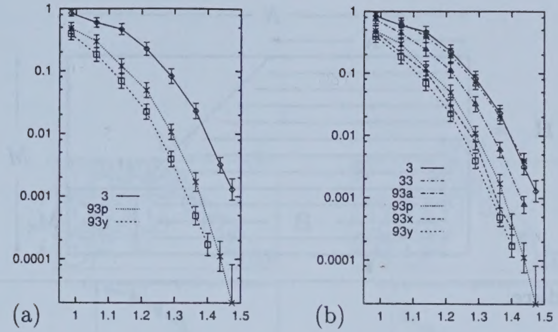
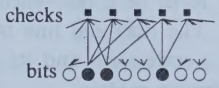


Figure 3: (a) Comparison of one representative of each of the constructions 3 (regular), 93p (Poisson) and 93y (super–Poisson). (b) shows representatives of all six constructions in figure 2. Vertical axis: block error probability. Horizontal axis: E_b/N_0 in dB.

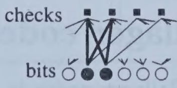
3.2.2 Explanation of error floors

In both cases the error floor has a simple cause. The most frequent error under these conditions is a reconstructed transmission which differs from the correct codeword in exactly three bits — the same three bits every time. These bits, which have weight 3 columns in the parity check matrix, are connected to just five checks with the topology shown below.



(1)

If the three bits shaded grey are flipped into the wrong state then the syndrome vector changes sign in the fifth check only. The sum–product algorithm is unable to extricate itself from this state. As the block length of the code is increased, the probability of this topology's occurrence falls. It is also possible to modify the construction algorithm for Gallager codes such that cycles of length 4, like this



(2)

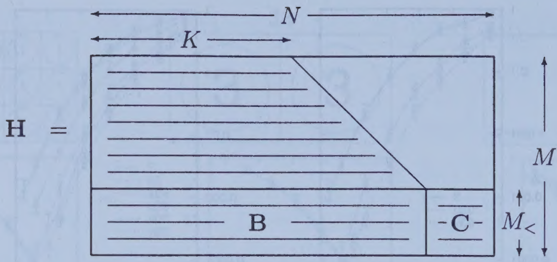
are forbidden (as in construction 1A of [9]). This modification is sufficient to prevent the topology shown in (1) from occuring.

We discard these two codes with error floors in the subsequent comparisons.

3.2.3 Comparison of constructions

The six families are compared with each other in figure 3. There are no detectable differences between the regular codes 3 and 33. There is a clear ranking of the other constructions, as follows:

$$3 < 93a < 93p < 93x < 93y$$



Encoding procedure:

Bits $t_1 \dots t_K$ are defined to be source bits.	
Bits $t_{K+1} \dots t_{N-M_<}$ are set in sequence, using the mth parity check to determine t_{K+m}.	This costs $(N - M_<)t_r$ computational operations, where t_r is the typical weight per row.
Bits $t_{N-M_<+1} \dots t_N$ are set equal to: $\mathbf{C}^{-1}\mathbf{B}\mathbf{t}^* \bmod 2$ where $\mathbf{t}^* = (t_1 \dots t_{N-M_<})^{\mathsf{T}}$ and $\mathbf{C}^{-1}$ is the inverse of $\mathbf{C}$ in modulo 2 arithmetic.	$\mathbf{C}^{-1}$ can be stored in $M_<^2$ bits of memory. The product $\mathbf{B}\mathbf{t}^*$ can be computed in $M_< t_r$ computational operations, and the multiplication by $\mathbf{C}^{-1}$ takes $M_<^2$ operations.

Figure 4: Upper panel: General form of a fast–encoding Gallager code. Horizontal stripes indicate low–weight rows. The diagonal line is a line of 1s. Lower panel: The fast encoding method to generate a codeword, and its computational cost, assuming an appropriate representation of the sparse matrix.

Thus we find that, at least for the 93 profile, sub–Poisson constructions are inferior to Poisson constructions, and super–Poisson constructions are significantly superior. In the case of 93y we see an improvement of about 0.05dB.

4 Fast–encoding Gallager codes

One of the possible drawbacks of Gallager codes is that their encoding time generally scales as N^2. Inspired by Spielman's [11] work, we have investigated constructions of Gallager codes whose profiles are similar to or identical to the 3 and 93 profiles above, but which are fast–encoding. The general form of parity check matrix for a fast–encoding Gallager code is shown in figure 4. The parity check matrix has an almost lower–triangular structure which allows all but a small number $M_<$ of the parity bits to be computed using sparse operations. The final $M_<$ parity bits can be computed in $M_<^2$ binary operations. If $M_<$ were as small as $\sqrt{M}$ then the codes would be linear–time encodeable.

4.1 Results

Construction 13 has profile identical to construction 3. Construction 193y has a profile identical to the 93 codes, and is most similar to 93y. These constructions are displayed

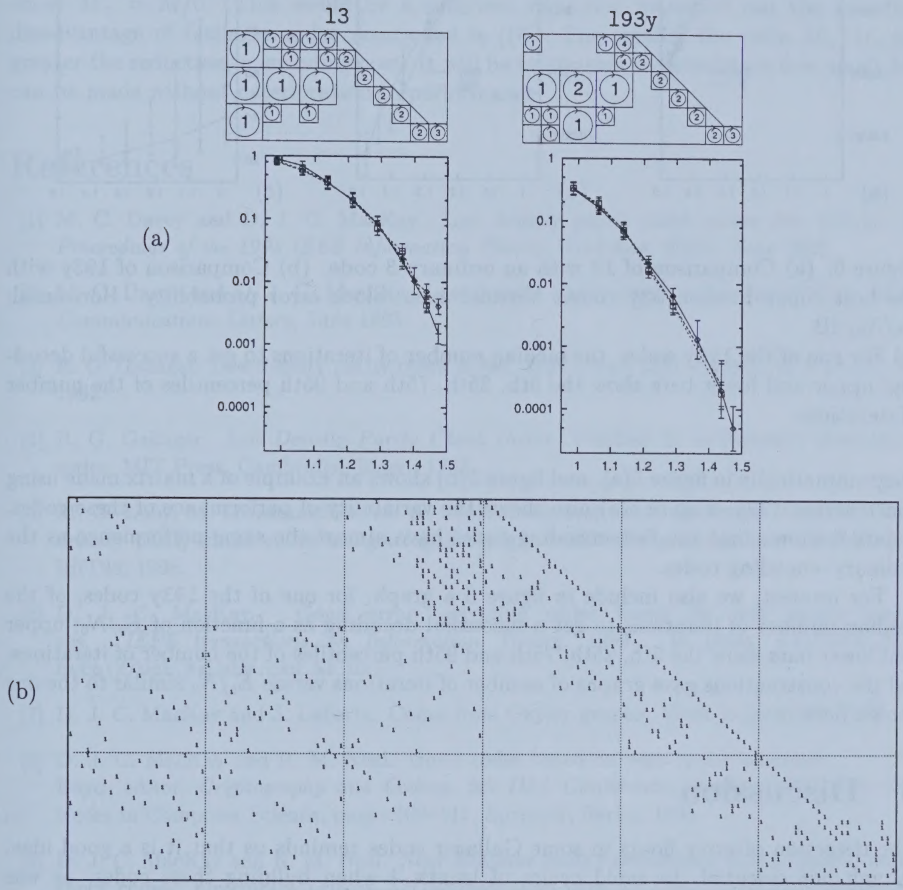


Figure 5: (a) Upper panels: construction methods 13 and 193y. As in figure 2, an integer represents a number of permutation matrices superposed on the surrounding square. A diagonal line represents an line of 1s. Horizontal and vertical lines indicate the boundaries of the permutation blocks. Lower pictures: Variability of performance among 13 and 193y codes. Vertical axis: block error probability. Horizontal axis: E_b/N_0 in dB. (b) Example of a parity check matrix with $N = 144$ made using construction 193y.

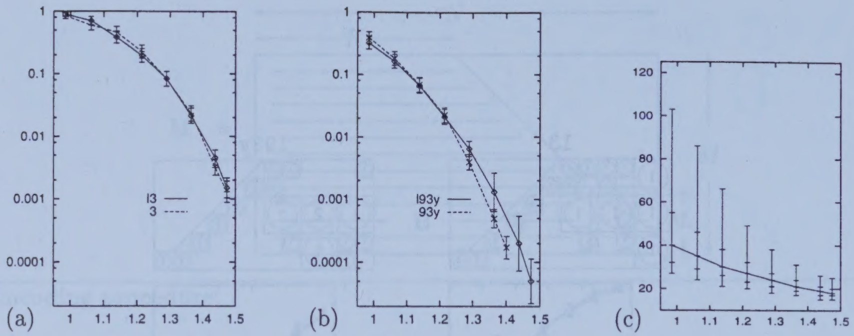


Figure 6: (a) Comparison of `13` with an ordinary `3` code. (b) Comparison of `193y` with the best super–Poisson `93y` code. Vertical axis: block error probability. Horizontal: E_b/N_0/dB.
(c) For one of the `193y` codes, the median number of iterations to get a successful decoding; upper and lower bars show the 5th, 25th, 75th and 95th percentiles of the number of iterations.

diagrammatically in figure 5(a), and figure 5(b) shows an example of a matrix made using construction `193y`. Figure 5(a) also shows the variability of performance of these codes. Figure 6 shows that the fast–encoding codes have almost the same performance as the ordinary–encoding codes.

For interest, we also include in figure 6 a graph, for one of the `193y` codes, of the median number of iterations to get a successful decoding as a function of E_b/N_0; upper and lower bars show the 5th, 25th, 75th and 95th percentiles of the number of iterations. All the constructions gave graphs of number of iterations versus E_b/N_0 similar to the one shown here.

5 Discussion

The detection of error floors in some Gallager codes reminds us that it is a good idea, though not essential, to avoid cycles of length 4 when building these codes, as was standard practice in [3, 9, 6]. This is not so easy to enforce in irregular Gallager codes with high weight columns, but the present results indicate that what is really important is to ensure that no pairs of low weight columns (with weights 2 or 3) have overlaps greater than one.

The improvement of 0.05dB of the super–Poisson construction compared with a Poisson construction of irregular codes may not be viewed as practically useful, but we think it is important to be aware of all the methods available for enhancing code performance, especially when these methods involve no added cost at the encoder or decoder. It will be interesting to see how beneficial the super–Poisson concept is for other irregular profiles. We are currently investigating the creation of irregular Turbo codes which make use of super–Poisson constructions (work in progress with B. J. Frey).

The discovery that we can make fast–encoding Gallager codes whose performance is just as good as ordinary Gallager codes may have immediate practical importance for mobile communications where computer power consumption must be minimized. Both the memory requirements and the CPU requirements at the encoder of our fast–encoding codes are substantially smaller. The particular constructions used in this paper would allow encoding using roughly thirty–six times fewer computational operations, since we chose $M_< = M/6$. This would be a sufficient reduction to cancel out the encoding disadvantage of Gallager codes mentioned in [10]. The smaller the ratio $M_</M$, the greater the reduction in encoding cost. It will be interesting to investigate how small $M_<$ can be made without deterioration in performance.

References

[1] M. C. Davey and D. J. C. MacKay. Low density parity check codes over GF(q). In *Proceedings of the 1998 IEEE Information Theory Workshop*. IEEE, June 1998.

[2] M. C. Davey and D. J. C. MacKay. Low density parity check codes over GF(q). *IEEE Communications Letters*, June 1998.

[3] R. G. Gallager. Low density parity check codes. *IRE Trans. Info. Theory*, IT-8:21–28, Jan 1962.

[4] R. G. Gallager. *Low Density Parity Check Codes.* Number 21 in Research monograph series. MIT Press, Cambridge, Mass., 1963.

[5] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman. Improved low–density parity–check codes using irregular graphs and belief propagation. Submitted to ISIT98, 1998.

[6] D. J. C. MacKay. Good error correcting codes based on very sparse matrices. *IEEE transactions on Information Theory*, 1999. In press. Available from `http://wol.ra.phy.cam.ac.uk/`.

[7] D. J. C. MacKay and J. Lafferty. Codes from Cayley graphs. Work in progress, 1997.

[8] D. J. C. MacKay and R. M. Neal. Good codes based on very sparse matrices. In Colin Boyd, editor, *Cryptography and Coding. 5th IMA Conference*, number 1025 in Lecture Notes in Computer Science, pages 100–111. Springer, Berlin, 1995.

[9] D. J. C. MacKay and R. M. Neal. Near Shannon limit performance of low density parity check codes. *Electronics Letters*, 32(18):1645–1646, August 1996. Reprinted *Electronics Letters*, **33**(6):457–458, March 1997.

[10] V. Sorokine, F. R. Kschischang, and S. Pasupathy. Gallager codes for CDMA applications II: Implementations, complexity and system capacity. *IEEE Trans. Communications*, 1998. submitted.

[11] D. A. Spielman. Linear-time encodable and decodable error-correcting codes. *IEEE Transactions on Information Theory*, 42(6.1):1723–1731, 1996.

Performance Analysis of Turbo-Equalized Dicode Partial-Response Channel

Mats Öberg and Paul H. Siegel

Dept. of Electrical and Computer Engineering
University of California, San Diego
La Jolla, California 92093-0407
{moberg, psiegel} @ucsd.edu

Abstract

This paper addresses the performance of maximum-likelihood decoding of a serial concatenation comprising a high-rate convolutional code or a turbo code, a uniform interleaver, and a dicode partial-response channel. The effect of a channel precoder on the system performance is also considered. Bit- and word-error-rate estimates based upon properties of the average Euclidean distance spectrum of the coded partial-response channel are derived. The estimates are compared to computer simulation results, and implications for system design are discussed.

1 Introduction

Trellis coding techniques that improve the reliability of binary input-constrained, intersymbol interference (ISI) channels are of interest in both digital communications and data storage applications. Drawing inspiration from the success of turbo codes [1], [2], several authors have recently considered iterative decoding architectures for coding schemes of the form depicted in Fig. 1, where the outer encoder is a convolutional or turbo encoder, π is an interleaver, $g(D)$ represents a precoder function, and $h(D)$ is the channel transfer polynomial.

This system resembles serial concatenation of interleaved codes, investigated by Benedetto, *et al.* [3], with the inner code replaced by the ISI channel. For such a system, Douillard, *et al.* [4] presented an iterative receiver structure, dubbed "turbo-equalization," to combat ISI due to multipath effects on convolutionally-coded Gaussian and Rayleigh transmission channels. They introduced an interleaver between the encoder and channel, and, as in turbo decoding, soft-output decisions from the channel detector and from the convolutional decoder are used in an iterative and cooperative fashion.

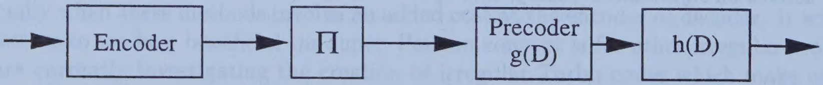


Figure 1: Trellis coded partial response system.

Motivated largely by the potential applications to digital magnetic recording, several authors have explored the application of turbo coding methods to the dicode and class-4 partial-response (PR4) channels, which have transfer functions $h(D) = 1 - D$ and $h(D) = 1 - D^2$, respectively. Heegard [5] and Pusch, *et al.* [6] illustrated the design and iterative decoding of turbo codes for the dicode channel, using rates 1/2 and lower. Reed and Schlegel [7], extending prior results on a low-complexity, iterative multiuser receiver structure with interference cancellation, have evaluated the benefits of turbo-equalization for a rate 1/2, convolutionally-coded, PR4 channel and E^2PR4 channel.

Ryan, *et al.* [8], and Ryan [9], demonstrated that by using as an outer code a parallel-concatenated turbo code, punctured to achieve rates 4/5, 8/9 and 16/17 typical of commercial magnetic recording systems, one could obtain significant coding gain relative to previously known high-rate trellis-coding techniques on a precoded dicode or PR4 channel.

Recently, Souvignier, *et al.* [10] considered serial concatenated systems similar to that addressed in [8],[9]. On a precoded dicode channel, the performance achievable with a high-rate outer convolutional code, rather than a turbo code, was investigated by means of computer simulation. Somewhat surprisingly, the convolutional code was found to perform as well as the turbo code. Moreover, removal of the channel precoder was found to improve the performance of the turbo-coded system at low SNR, while degrading the performance of the convolutionally-coded system.

This paper was motivated, in part, by the desire to understand better the empirically observed differences in error-rate in the precoded and non-precoded serial concatenated systems. We address the performance of maximum-likelihood decoding of a serial concatenated system as shown in Fig. 1, comprising a high-rate convolutional encoder or a parallel concatenation of two convolutional encoders, an interleaver, and a dicode partial-response channel, with and without a channel precoder.

The maximum–likelihood (ML) union bound on word error rate (WER) for a block-coded, additive white Gaussian noise (AWGN) channel can be expressed as [11]

$$P_w \leq \sum_{d_E=d_{min}}^{\infty} \overline{T}(d_E) Q\left(\frac{d_E}{2\sigma}\right), \tag{1}$$

where d_E denotes Euclidean distance between two channel output words, $\overline{T}(d_E)$ denotes the average Euclidean weight enumerator which is the average number of codewords with Euclidean distance d_E from the output of a given codeword, and σ^2 denotes the noise variance on the channel. The corresponding bit error rate (BER) bound is [11]

$$P_b \leq \sum_{d_E=d_{min}}^{\infty} \frac{\overline{T}(d_E)\overline{w}(d_E)}{K} Q\left(\frac{d_E}{2\sigma}\right), \tag{2}$$

where $\overline{w}(d_E)$ denotes the average information Hamming distance to codewords whose outputs have distance d_E, and K denotes the number of information bits in a codeword.

For an exact analysis, the full compound error-event characterization for a code interleaved and concatenated with the ISI channel must be determined. The complexity of this computation is generally prohibitively high. To overcome this difficulty, we introduce a technique for computing an approximation to the average weight enumerator $\overline{T}(d_E)$ for a high-rate, coded partial response channel. The result depends only upon the output Hamming weight enumerator of the outer code,

$$\overline{A}(d) = \sum_{i=0}^{K} \overline{A}(d, i), \tag{3}$$

where $\overline{A}(d, i)$ denotes the average number of error words of Hamming output weight d and input weight i.

In Section 2, we present the analysis for the dicode channel, first without a precoder, then with a precoder of the form $g(D) = 1/(1 \oplus D)$. In Section 3, we consider systems incorporating a rate 8/9 outer punctured convolutional code and a rate 4/5 turbo code. The analytical performance estimates are compared to the results of computer simulation. Section 4 concludes the paper.

2 Error Event Analysis on the Dicode Channel

Referring to the system model in Fig. 1, we assume that the encoder is a block encoder, for example, a truncated convolutional encoder or a turbo encoder. Let $\mathbf{b} = [b_1, \ldots, b_N]$ denote a codeword, and $\mathbf{c} = \pi(\mathbf{b})$ be the corresponding output of the interleaver. The output of the precoded channel is denoted $\mathbf{x} = [x_1, \ldots, x_N]$.

Given two codewords $\mathbf{b_1}$ and $\mathbf{b_2}$, let $\mathbf{e} = \mathbf{b_1} \oplus \mathbf{b_2}$ be the corresponding Hamming error word, and similarly define the interleaved Hamming error word $\mathbf{f} = \mathbf{c_1} \oplus \mathbf{c_2}$. The Euclidean error word is given by $\zeta = \mathbf{c_1} - \mathbf{c_2}$ and the corresponding output error word is $\chi = \mathbf{x_1} - \mathbf{x_2}$.

We will make two simplifying assumptions in the analysis of the system performance. First, we assume that the interleaver π is a uniform interleaver, as defined by Benedetto et al. [12]. Second, we make the assumption that, for any codeword $\mathbf{e}$, the contribution to $\overline{T}(d_E)$ of all error words $\epsilon = \mathbf{b_1} - \mathbf{b_2}$, where $\mathbf{b_1} = \mathbf{b_2} \oplus \mathbf{e}$, is approximately equal to the contribution of the set of error words produced when $\mathbf{b_1}$ and $\mathbf{b_2}$ are not restricted to lie within the code. This is equivalent to treating the permuted code bits within an error event at the output of the interleaver as independent and identically distributed (i.i.d.), with equiprobable bit values. The resulting estimate of the contribution to the Euclidean weight enumerator therefore depends only upon the Hamming weight of the word $\mathbf{e}$. The rationale behind this second assumption is that the system uses a very high-rate linear encoder, in tandem with the uniform interleaver.

In Section 2.1, we investigate the relationship between the Hamming weight $d_H(\mathbf{f})$ of the interleaved Hamming error word and the squared Euclidean distance $d_E^2(\zeta) = \sum_{i=1}^{N} {\chi_i}^2$ of the corresponding output error word on the dicode channel, $h(D) = 1 - D$. We then examine the distribution of error events induced by the action of the uniform interleaver. Using these results, we then derive an estimate for the system performance.

In Section 2.2, we derive the corresponding result for the dicode channel with a precoder characterized by $g(D) = 1/(1 \oplus D)$.

2.1 Dicode Channel With No Precoder

2.1.1 Error Event Distance Properties

Fig. 2 shows a trellis section for the dicode channel with no precoder. The branch labels are of the form c_i/x_i, where c_i is the input to the channel at time i, and x_i is the corresponding channel output.

Let f be an error word with Hamming weight $l = d_H(f)$, corresponding to a possibly compound input error event. Referring to Fig. 2, and assuming a fixed initial state for all codewords, f can be uniquely decomposed into a concatenation of disjoint error sub-events f_i, $i = 1, \ldots, m$, for some $m \geq 1$, consisting of one or more consecutive errors.

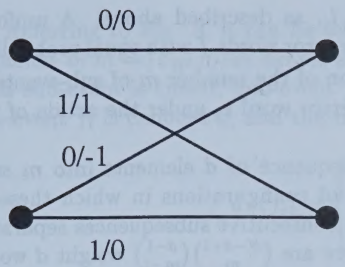


Figure 2: Trellis section for the dicode channel.

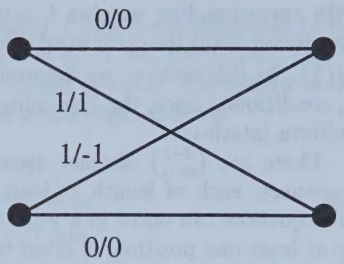


Figure 3: Trellis section for the precoded dicode channel.

Letting $l_i = d_H(f_i)$ denote the number of errors in the sub-event, we have $l = \sum_{i=1}^{m} l_i$. For $i \leq m-1$, the sub-event f_i corresponds to a simple closed error event on the trellis, diverging from and remerging with the correct path, with no common intermediate states. For $i = m$, the corresponding sub-event f_m may be either closed or open; in the latter case the paths diverge and never remerge.

Let j_i denote the bit position at which the sub-event f_i begins. For a closed sub-event, the contribution to the squared Euclidean distance is given by

$$d_E^2(f_i) = 2 + 4 \sum_{k=j_i+1}^{j_i+l_i-1} c_k \oplus c_{k-1}. \tag{4}$$

If f_m is open, the contribution is

$$d_E^2(f_m) = 1 + 4 \sum_{k=j_m+1}^{j_m+l_m-1} c_k \oplus c_{k-1}. \tag{5}$$

The compound error event f generates squared Euclidean distance

$$d_E^2(f) = \sum_{i=1}^{m} d_E^2(f_i) = 2m + 4 \sum_{i=1}^{m} \sum_{k=j_i+1}^{j_m+l_m-1} c_k \oplus c_{k-1} - \delta(j_m + l_m - 1 - N), \tag{6}$$

where

$$\delta(n) = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Invoking the assumption regarding the distribution of code bit values in the error events – namely, that their values are i.i.d. and equiprobable – we obtain an approximate contribution of an error word f of Hamming weight d to the average Euclidean distance spectrum. For example, for the case when f_m is closed, the result is:

$$Pr(d_E^2(f) = 2m + 4y | d, m, f_m \text{ closed}) = \binom{d-m}{y} 0.5^{d-m}. \tag{8}$$

The derivation of the approximate contribution for the case where f_m is open is similar.

2.1.2 Sub-Event Distribution at the Interleaver Output

Let e be an error word with Hamming weight $d_H(e) = d$. A specified interleaver will map e into an error word f which can be decomposed into m error sub-events, f_i, $i = 1, \dots, m$, with corresponding weights l_i satisfying $l = \sum_{i=1}^{m} l_i$, as described above. A uniform interleaver maps the error word e into all $\binom{N}{d}$ possible error words f with equal probability $1/\binom{N}{d}$. In this section, we determine the distribution of the number m of sub-events of f, conditioned upon the Hamming weight d of the error word e, under the action of the uniform interleaver.

There are $\binom{d-1}{m-1}$ distinct decompositions of a sequence of d elements into m subsequences, each of length at least 1. The number of configurations in which these m subsequences can occur in a word of length N, with consecutive subsequences separated by at least one position, is given by $\binom{N-d+1}{m}$, so there are $\binom{N-d+1}{m}\binom{d-1}{m-1}$ weight d words with m sub-events. Taking into consideration the nature of the sub-event f_m, we can compute the conditional joint probabilities:

$$Pr(m, f_m \; closed|d) = \frac{\binom{N-d}{m}\binom{d-1}{m-1}}{\binom{N}{d}} \tag{9}$$

and

$$Pr(m, f_m \; open|d) = \frac{\binom{N-d}{m-1}\binom{d-1}{m-1}}{\binom{N}{d}}. \tag{10}$$

2.1.3 Approximation of the Euclidean Weight Enumerator

The approximation $\tau(d_E)$ to the average Euclidean weight enumerator $\overline{T}(d_E)$ can be computed by substituting (8) and its counterpart for the open case, along with (9) and (10) into

$$\begin{aligned} \overline{T}(d_E) &= \sum_{k=1}^{N} \overline{A}(k) Pr\left(d_E | d_H = k\right) \\ &= \sum_{k=1}^{N} \overline{A}(k) \sum_{m=1}^{k} \left(Pr\left(d_E | d_H = k, m, f_m \; closed\right) Pr\left(m, f_m \; closed | d_H = k\right)\right. \\ &\quad + \left. Pr\left(d_E | d_H = k, m, f_m \; open\right) Pr\left(m, f_m \; open | d_H = k\right)\right). \end{aligned} \tag{11}$$

The approximate average input error weight enumerator is similarly computed by substitution into

$$\overline{w}(d_E) = \frac{1}{\overline{T}(d_E)} \sum_{k=1}^{N} \overline{A}(k) \overline{W}(k) \sum_{m=1}^{k} Pr\left(d_E | d_H = k, m\right) Pr\left(m | d_H = k\right), \tag{12}$$

where $\overline{W}(k)$ is the average input weight for codewords of Hamming weight $d_H = k$.

2.2 Precoded Dicode

2.2.1 Error Event Distance Properties

Fig. 3 shows a trellis section for the dicode channel with precoder $g(D) = 1/(1 \oplus D)$. The branch labels are of the form c_i/x_i, where c_i is the input to the precoder at time i, and x_i is the corresponding channel output.

Referring to Fig. 3, it can be seen that an error word f may be decomposed into a sequence of $m = \lceil d_H(f)/2 \rceil$ simple error sub-events f_i, $i = 1, \dots, m$. For $1 \le i \le m-1$, each sub-event is closed, Sub-event f_m may be either closed or open. The length of the sub-event f_i is denoted l_i, and the Hamming weight of a sub-event satisfies

$$d_H(f_i) = \begin{cases} 2 & i = 1, \dots, m-1 \\ 2 & i = m \text{ and } d_H(f) \text{ even} \\ 1 & i = m \text{ and } d_H(f) \text{ odd.} \end{cases} \tag{13}$$

Let j_i^0 denote the bit position in the word where error event f_i begins. For closed events, let j_i^1 denote the bit position where it terminates. Then $l_i = j_i^1 - j_i^0 + 1$ for all closed sub-events, If f_m is open, we define $j_m^1 = N + 1$, and $l_m = j_m^1 - j_m^0$. Finally we define $L = \sum_{i=1}^m l_i$.

For an error event f_i, the total contribution $d_E^2(f_i)$ to the squared Euclidean distance at the channel output is given by

$$d_E^2(f_i) = d_H(f_i) + 4 \sum_{k=j_i^0+1}^{j_i^1-1} c_k. \tag{14}$$

The error word f has total squared Euclidean distance

$$d_E^2(f) = \sum_{i=1}^{m} d_E^2(f_i) = d_H(f) + 4 \sum_{i=1}^{m} \sum_{k=j_i^0+1}^{j_i^1-1} c_k. \tag{15}$$

We now invoke the assumption regarding the distribution of code bit values in the error events – namely, that their values are i.i.d. and equiprobable – to obtain an approximate contribution of an error word f of Hamming weight d to the average Euclidean distance spectrum. Under this assumption, we obtain

$$Pr(d_E^2(f) = z | d_H(f) = d, L) = \binom{L-d}{(z-d)/4} 0.5^{L-d}. \tag{16}$$

The i.i.d. approximation is supported for error words **f** with a small value of L by the action of the uniform interleaver. For error words with large value of L, the contribution to the dominant terms of the Euclidean error weight enumerator will be negligible, due to the low probability of generating small Euclidean distance.

2.2.2 Sub-Event Distribution at the Interleaver Output

Let **e** be an error word of Hamming weight d. A permuted error word **f** can be decomposed into $m = \lceil d/2 \rceil$ error events f_i, as described in Section 2.2.1. In this section, we determine the distribution of the total length L of error words generated by the action of a uniform interleaver upon the error word **e**.

The distribution is computed in two steps. First, we find the number of unique back-to-back concatenations of m sub-events of total length L. Then, we determine the number of configurations in which the m sub-events can occur in a word of length N.

Consider the following description of the permuted error word f,

$$0, \dots, 0, 1_1, 0_1, \dots, 0_1, 1_1, 0, \dots, 0, \dots, 1_m, 0_m, \dots, 0_m, 1_m, 0, \dots, 0,$$

where the subscript denotes to which sub-event a bit belongs. There are $\binom{L-\lceil (d-1)/2\rceil-1}{\lfloor (d-1)/2\rfloor}$ unique back-to-back concatenations of sub-events f_i of total length L. If d is even, the remaining $N-L$ bits can be partitioned in $\binom{N-L+m}{m}$ different ways. If d is odd, the permutation has to end with an open error event, so there are $\binom{N-L+m-1}{m-1}$ possible permutations.

The conditional distribution $Pr(L|d)$ of the total length L, given an error word of Hamming weight $d_H(e)=d$, is therefore given by

$$Pr(L|d) = \frac{\binom{N-L+\lfloor d/2\rfloor}{\lfloor d/2\rfloor}\binom{L-1-\lceil (d-1)/2\rceil}{\lfloor (d-1)/2\rfloor}}{\binom{N}{d}}. \tag{17}$$

2.2.3 Approximation of the Euclidean Weight Enumerator

The approximate Euclidean weight enumerator $\tau(d_E)$ can be computed by substituting (16) and (17) into

$$\overline{T}(d_E) = \sum_{k=1}^{N} \overline{A}(k) \sum_{L=k}^{N-k} Pr\left(d_E|k,L\right) Pr\left(L|k\right). \tag{18}$$

In a similar way, the approximate average input error weight enumerator may be obtained from

$$\overline{w}(d_E) = \frac{1}{\overline{T}(d_E)} \sum_{k=1}^{N} \overline{A}(k)\overline{W}(k) \sum_{L=k}^{N-k} Pr\left(d_E|k,L\right) Pr\left(L|k\right). \tag{19}$$

3 Computed Bounds and Simulation Results

In this section, we compute truncated maximum likelihood (ML) union bound estimates using the method outlined in the paper, and we compare these with computer simulation results. We consider two outer encoders: (1) a rate 1/2, recursive systematic convolutional (RSC) encoder with encoder polynomials $(31,33)_{octal}$, with parity bits punctured to yield code rate 8/9; and (2) a rate 4/5 turbo code consisting of a parallel concatenation of two of the RSC encoders. Both encoders use an information block of size $K=4096$. The decoding technique used in the simulations is iterative, with *a posteriori* probability (APP) decoders used for the channel detectors as well as the decoders. Soft information is shared between all decoders for ten full iterations.

Fig. 4 shows the word error rate (WER) results for the rate 8/9 system, with and without precoder. Fig. 5 shows the corresponding results for bit error rate (BER).

In Table 1, the Euclidean weight enumerator estimates for the two systems are shown. (The authors are thankful to Dr. Douglas N. Rowitch [13] for providing us with weight enumerators.) For the dicode channel, the dominant contributor to the estimated error rate is the $d_E^2=4$ component; with the precoder, it is the $d_E^2=2$ component.

The interleavers used in the simulations and the uniform interleaver induce different weight enumerators. Therefore the estimated bounds and the simulation results differ and in the case without precoder the simulation curve crosses the bound curve. Fig. 6 compares the BER simulation results with the estimates obtained by applying our analytical approach to the specific interleaver. The fit between the analysis and simulation is improved, particularly in the precoded case. Without the precoder, accounting for the

Interleaver	1	2	3	4
BER (6.0 dB) $/10^{-5}$	0.6610	0.2689	0.0766	0.0407

Table 2: BER with precoder at $E_b/N_0 = 6.0$ dB for three interleavers.

Outer Code			Precoded	Not Precoded
d_H	$\overline{A}(d_H)$	d_E^2	$\tau(d_E)$	$\tau(d_E)$
2	0.031	2	0.000024	0.000006
3	0.462	3	0.000000	0.000000
4	2.111	4	0.000004	0.031273
5	4.100	5	0.000000	0.000000
6	8.842	6	0.000024	0.464084
7	20.337	7	0.000001	0.000000
8	50.743	8	0.000008	2.114827

Table 3: Approximate Hamming and Euclidean weight enumerators for turbo coded systems.

specific interleaver used in the simulation made only a minor difference in the dominant term $\overline{T}(d_E^2 = 4)$. For smaller Euclidean distances, the relative changes were larger; for example, for the interleaver used in the simulations, the estimated weight enumerator for $d_E^2 = 2$ was 10 times larger than the average.

In contrast, with the precoder, the interleaver choice has a significant impact upon the estimate of the dominant term $\overline{T}(d_E^2 = 2)$. For the interleaver used, we computed the contribution made to the estimated Euclidean weight enumerator by error events of length $L \leq 15$ corresponding to error words of Hamming weight $d_H = 2$. The result was approximately 10 times the contribution of the average interleaver, leading to substantial improvement in the accuracy of the estimated error rate.

The impact of the interleaver in the precoded case is further reflected in simulation results for three additional interleavers. Table 2 shows their BER values at $E_b/N_0 = 6.0$ dB. The table suggests that suitable interleaver design can significantly improve the system performance.

Fig. 7 shows analytical BER estimates and simulation results for the rate 4/5 turbo coded systems. The Hamming weight enumerator and the estimated Euclidean weight enumerator for the turbo-coded system are shown in Table 3.

The bound for the precoded system is much lower at SNR up to about $E_B/N_0 = 9.5$ dB. However, in simulations, the system without precoder is superior down to $P_b \approx 2 \cdot 10^{-7}$, at which point the simulated BER curve flattens out and tends to follow the analytical curve. In fact, above 4.7 dB, the precoded system becomes superior to system without the precoder, as predicted by the analysis. The explanation for the behavior observed at very low SNR remains an open issue.

4 Conclusions

We have presented an analytical method for estimating the average Euclidean distance spectrum for a serial concatenated, trellis-coded partial response channel. The technique was applied to the dicode channel, with and without precoding. Using truncated union bounds, we derived analytical BER and WER results and compared them to computer simulations.

Future research directions are to bound the effect of the i.i.d. assumption, develop methods for higher order channels, and include the entire Hamming distance spectrum in the computations.

References

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo Codes," in *Proc. IEEE Int. Conf. Commun.*, (Geneva, Switzerland), pp. 1064–1070, May 1993.

[2] C. Berrou and A. Glavieux, "Near optimum error-correcting coding and decoding: Turbo Codes," *IEEE Trans. Commun.*, vol. 44, pp. 1261–1271, Oct. 1996.

[3] S. Benedetto, G. Montorsi, D. Divsalar, and F. Pollara, "Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding," *The Telecom. and Data Acquisition Progress Report*, vol. 42, pp. 1–26, Aug. 1996.

[4] C. Douillard, M. Jézéquel, C. Berrou, A. Picart, P. Didier, and A. Galieux, "Iterative correction of intersymbol interference: Turbo-equalization," *European Trans. Telecommunications*, vol. 6, pp. 507–511, Sept./Oct. 1995.

[5] C. Heegard, "Turbo coding for magnetic recording," in *Proceedings Winter 1998 Information Theory Workshop*, (San Diego, CA), pp. 18–19, IEEE, Feb. 1998.

[6] W. Pusch, D. Weinrichter, and M. Taferner, "Turbo-codes matched to the $1 - d^2$ partial response channel," in *Proc. of IEEE Int. Symposium on Inform Theory*, (Cambridge, MA), p. 62, Aug. 1998.

[7] M. C. Reed and C. B. Schlegel, "An iterative receiver for the partial response channel," in *Proc. of IEEE Int. Symposium on Inform Theory*, (Cambridge, MA), p. 63, Aug. 1998.

[8] W. E. Ryan, L. L. McPheters, and S. W. McLaughlin, "Combined turbo coding and turbo equalization for PR4-equalized Lorentzian channels," in *Proc. of the Conf. on Inform Sciences and Systems*, March 1998.

[9] W. E. Ryan, "Performance of high rate turbo codes on a PR4-equalized magnetic recording channel," in *Proc. IEEE Int. Conf. Commun.*, (Atlanta, GA), June 1998.

[10] T. Souvignier, A. Friedman, M. Öberg, P. H. Siegel, R. E. Swanson, and J. K. Wolf, "Turbo codes for PR4: Parallel versus serial concatenation." Submitted to ICC '99.

[11] G. D. Forney Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 363–378, May 1972.

[12] S. Benedetto and G. Montorsi, "Unveiling turbo codes: Some results on parallell concatenated coding schemes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 409–428, March 1996.

[13] D. N. Rowitch, *Convolutional and Turbo Coded Multicarrier Direct Sequence CDMA, and Applications of Turbo Codes to Hybrid ARQ Communication Systems.* PhD thesis, University of California San Diego, La Jolla CA 92093-0407, June 1998.

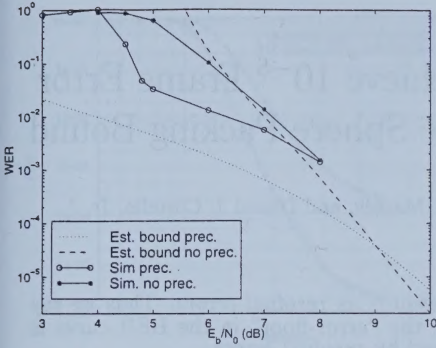


Figure 4: Word error rate union bound estimates and simulation results.

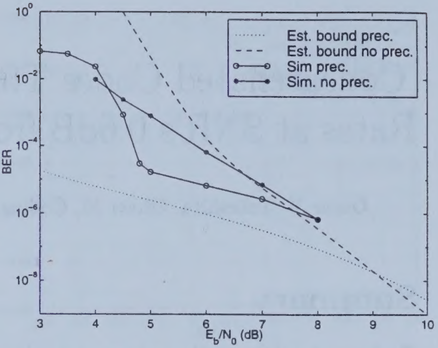


Figure 5: Bit error rate union bound estimates and simulation results.

Outer Code			Precoded	Not Precoded
d_H	$\overline{A}(d_H)$	d_E^2	$\tau(d_E)$	$\tau(d_E)$
2	510	2	0.4426	0.1122
3	21421	3	0.02421	0.2274
4	357864	4	0.8084	523.6
5	13192299	5	0.06468	14.25
6	389079383	6	4.255	21864
7	9010184299	7	0.3166	336.1
8	236369355044	8	15.68	386587

Table 1: Hamming and approximate Euclidean weight enumerators for systems with outer convolutional code.

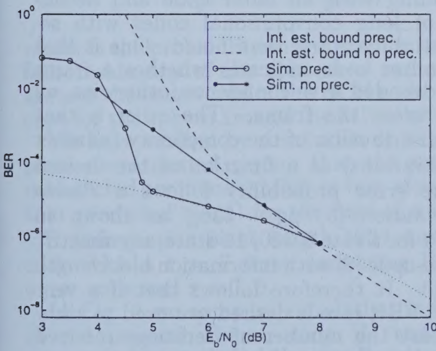


Figure 6: Truncated bound estimates for interleaver used vs. simulation result.

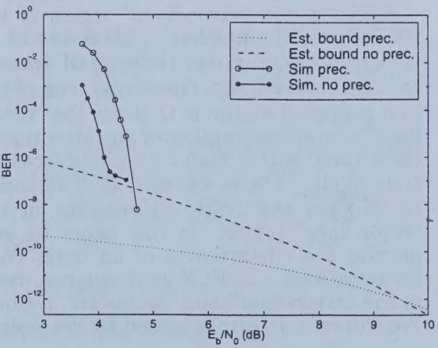


Figure 7: Bit error rate bound estimates vs. simulation results for precoded and non-precoded turbo system.

Concatenated Codes That Achieve 10^{-6} Frame Error Rates at SNR's 0.6dB from the Sphere Packing Bound

Oscar Y. Takeshita, Oliver M. Collins, Peter C. Massey, and Daniel J. Costello, Jr. [*]

Summary

Turbo-codes [1] with long frame lengths are usually constructed using a random interleaver. Statistically, this guarantees excellent bit error rate (BER) performance but also generates a small number of low weight codewords, resulting in the appearance of an "error-floor" at BER's below 10^{-5}. Several methods, including using an outer code, have been proposed to improve the "error-floor" region of the BER curve. When the codelength is known, [2] argues that (frame error rate) FER is a better performance measure for a code than BER and shows that turbo-codes are about 0.7dB away from the sphere packing bound for several frame lengths and code rates at a FER of 10^{-4}. In this paper we study the effect of an outer BCH code on the FER of turbo-codes. We show that additional coding gain is possible not only in the "error-floor" region but also in the "waterfall" region. With this method, we obtain codes whose performance is within 0.6dB of the sphere packing bound at an FER of 10^{-6}.

Since, in the "error-floor" region of the BER curve, the number of bit errors in an incorrect frame is relatively small (caused by a few low weight codewords), one of the first proposed methods to lower the "error-floor" was concatenation of an entire turbo-code frame with a high rate algebraic outer code [3][4]. These schemes were intended to improve the BER performance in the "error-floor" region. In this paper we emphasize the effectiveness of an outer code for improving the FER performance, both in the "error-floor" and "waterfall" regions. We will refer to errors caused by low weight codewords as *residual errors.* Thus we say that the "error-floor" in the BER curve is caused by *residual errors.*

We have further observed that for some turbo-codes *residual errors* are also present in large proportion in the "waterfall" region of the BER curve. Although the contribution of the *residual errors* is not significant for the BER curve in the "waterfall" region, it can significantly affect the FER curve. Therefore a scheme that removes the *residual errors* can greatly improve the FER performance of turbo-codes operating at low to moderate SNR's, i.e., in the "waterfall" region.

In addition to improving the BER and FER performance of turbo-codes, there are several other advantages of using a BCH outer code over an entire frame. One is that it provides a reliable stopping criterion for the iterative MAP decoder. Indeed, simulation results indicate that some turbo-codes must operate with a reliable stopping criterion in order to achieve good BER performance.

Interesting similarities between turbo-decoding with an outer code and decoding of long convolutional codes with sequential decoding were noted. One is that the outer code indicates whether a frame was decoded with confidence; otherwise, we may erase the frame. The other is that the distribution of the complexity (number of iterations) as a function of the desired frame error probability follows a *Pareto distribution* [5, pg. 336], as shown in Fig. 1 for a rate 0.330, 16-state, asymmetric turbo-code [6] with information blocklength 16243. It therefore follows that if a very small FER rate is desired, we need to allow at least the number of iterations inferred from the *Pareto distribution* and then remove the *residual errors* with a BCH code.

Consider a BCH code with parameters (n, k, t, τ), where n is the block length, k is the information length, t is the design error correction capability, and $\tau \leq t$ is the maximum number of errors actually corrected,

---

[*] This work was supported by NSF Grants NCR95-22939 and NCR96-96065 and NASA Grant NAG5-557. Oscar Y. Takeshita, Oliver M. Collins, Peter C. Massey, and Daniel J. Costello, Jr. are with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46656.

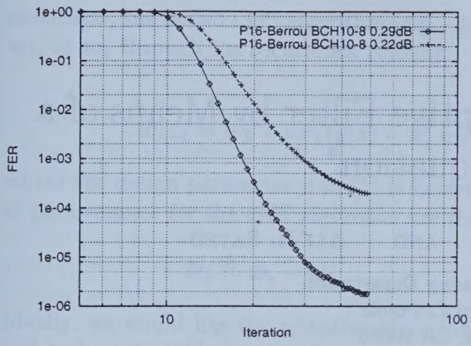


Figure 1: FER in function of the maximum number of iterations.

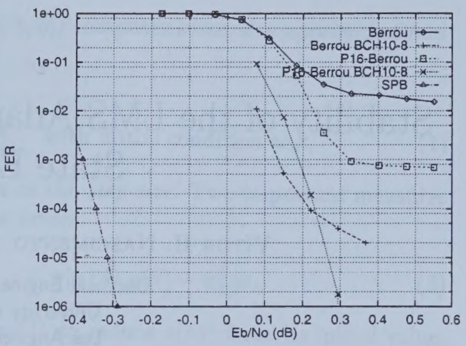


Figure 2: FER of turbo-codes with and without concatenation.

i.e., up to $2(t-\tau)$ additional errors can be detected, resulting in a decoding failure. If we assume that the decisions made by the iterative MAP decoder are uncorrelated to the BCH codewords after each iteration, except for the case when the iterative MAP decoder generates a codeword whose distance from the ML solution is smaller than the error correcting capability of the BCH code, then the undetected error probability, i.e., the probability that an (n,k,t,τ) BCH code falsely stops after a given iteration of the iterative MAP decoder, is given by:

$$P_u = \frac{(2^k-1)\binom{n}{\tau}}{2^n} \approx \frac{\binom{n}{\tau}}{2^{(n-k)}}, \qquad (1)$$

which is the probability that a random word falls inside any of the spheres of radius τ around the codewords of the BCH code, except for the sphere around the ML solution.

We have verified by simulation of a complete concatenated system that the actual undetected error probability is very close to that predicted by (1) for the case of a rate 1/3, information block length 1024, turbo-code with a $t=7$ error correcting BCH outer code and $\tau = 7, 6$.

In Fig. 2 we compare simulation results of the original Berrou code [1] and an asymmetric code [6] without concatenation (rate 1/3, information blocklength 16384 bits, fixed number of iterations=18) and with concatenation (rate 0.330, information blocklength 16243, $t=10$, $\tau=8$, maximum of 50 iterations, average of 11.5 iterations at 0.29dB). The effectiveness of concatenation can be seen from the simulations, where at a frame error rate of 10^{-6} the new codes are within 0.6dB of the sphere packing bound, which equals -0.288dB for rate 0.330 and blocklength 16243.

References

[1] C.Berrou et.al., “Near Shannon limit error-correcting coding and decoding: turbo-codes,” *ICC Proceedings,* pp. 1064-1070, May 1993.

[2] S.Dolinar et.al., “Code performance as a function of block size,” TMO Progress Report 42-133, May 1998.

[3] J.D.Andersen, “Turbo codes extended with outer BCH code,” *Electronics Letter,* vol. 32, no. 22, pp. 2059-2060, Oct. 1996.

[4] D.J.Costello, Jr. and G.Meyerhans, “Concatenated turbo codes,” *Proc. 1996 IEEE Intnl. Symp. on Inform. Theory and Its Applic.,* pp. 571-574, Sept. 1996.

[5] S.Lin and D.J.Costello, *Error control coding: fundamentals and applications,* Prentice-Hall, 1983.

[6] O.Y.Takeshita et.al., “Asymmetric Turbo-Codes,” *Proc. 1998 IEEE Intnl. Symp. on Inform. Theory,* pp. 179, Aug. 1998.

Stability of the LMS Adaptive Filter by Means of a State Equation*

VÍTOR H. NASCIMENTO AND ALI H. SAYED

Electrical Engineering Department
University of California
Los Angeles, CA 90095

Abstract

This work studies the mean-square stability of stochastic gradient algorithms without resorting to slow adaptation approximations or to the widely used, yet rarely applicable, independence assumptions. This is achieved by reducing the study of the mean-square convergence of an adaptive filter to the study of the exponential stability of a linear time-invariant state equation. The size of the coefficient matrix of the state equation, however, turns out to grow exponentially fast with the length of the filter so that it becomes computationally infeasible to manipulate the matrix directly. It is instead shown that the coefficient matrix is sparse and has structure. By exploiting these two properties, and by applying a sequence of carefully chosen similarity transformations to the coefficient matrix, an upper bound on the step-size is found that guarantees stability.

1 Introduction

Consider the following estimation problem. Given a zero-mean scalar sequence $\{y(k)\}_{k=0}^{\infty}$ and another sequence of zero-mean length-M column vectors $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$, we seek a length-$M$ column vector $\{\boldsymbol{w}_*\}$ that solves

$$\min_{\boldsymbol{w}} \mathrm{E}\big(y(k) - \boldsymbol{x}_k^T \boldsymbol{w}\big)^2.$$

where E is the expectation operator, k is the time index, and ^{T} is the transposition operator. The sequence $\{y(k)\}$ is called the *desired* sequence and $\{\boldsymbol{x}_k\}$ is the *input* or *regressor* sequence; the sequences $\{y(k), \boldsymbol{x}(k)\}$ are assumed to be jointly stationary. The error sequence is defined by $v(k) = y(k) - \boldsymbol{x}_k^T \boldsymbol{w}_*$ and its variance (also the minimum cost) is denoted by σ_v^2; we also refer to $\{v(k)\}$ as the *noise* sequence.

The optimal solution $\boldsymbol{w}_*$ can be fully characterized in terms of the second-order statistics of the random processes $\{y(k), \boldsymbol{x}(k)\}$. In adaptive implementations of the estimator $\boldsymbol{w}_*$, however, these statistics are often replaced by instantaneous approximations. In

---

*This material was based on work supported in part by the National Science Foundation under awards MIP-9796147 and CCR-9732376. The work of V. H. Nascimento was also supported by a fellowship from Conselho Nacional de Pesquisa e Desenvolvimento - CNPq - Brazil, while on leave from LPS-DEE, Escola Politécnica da Universidade de São Paulo, Brazil.

particular, the famed LMS algorithm (e.g., [1, 2, 3]) computes successive approximations, $\boldsymbol{w}_k$, to $\boldsymbol{w}_*$ through the recurrence relation

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \mu \boldsymbol{x}_k \big(y(k) - \boldsymbol{x}_k^T \boldsymbol{w}_k\big), \quad \text{with initial condition } \boldsymbol{w}_0 \;, \tag{1}$$

where the design parameter μ in (1) is known as the *step-size.* Two important measures of performance for the LMS algorithm are the errors

$$\tilde{\boldsymbol{w}}_k \triangleq \boldsymbol{w}_* - \boldsymbol{w}_k, \qquad\qquad e(k) \triangleq y(k) - \boldsymbol{x}_k^T \boldsymbol{w}_k. \tag{2}$$

Ideally, we would like the adaptive algorithm (1) to reduce $e(k)^2$ from its initial value, and to keep $\mathrm{E}\, e(k)^2$ close to σ_v^2 in steady-state (or, equivalently, $\tilde{\boldsymbol{w}}_k$ close to $\mathbf{0}$).

1.1 Performance with Slow Adaptation

The performance of an adaptive filter is often measured in terms of the variance $\mathrm{E}\, e(k)^2$, also known as the *mean-square error* or MSE for short, and by the trace of the covariance matrix of $\tilde{\boldsymbol{w}}_k$, $\mathrm{Tr}(\mathrm{E}\, \tilde{\boldsymbol{w}}_k \tilde{\boldsymbol{w}}_k^T)$, also called the *mean-square deviation* or MSD. Exact expressions for these measures can be hard to obtain in a general setting, which is in part due to the fact that adaptive systems are naturally time-variant and nonlinear. Nevertheless, there is a vast literature on the analysis of the performance of the LMS algorithm and its variants, which have been motivated by the large range of applications in which these algorithms have been successful. In most of these analyses, simplifying assumptions on the data and the model are imposed in order to make the derivation and the results more tractable.

The most important of these simplifying conditions are known collectively as the *independence assumptions*, which appear in many of the earlier works on LMS and in several recent ones (see, e.g., [1, 4, 5, 6]). Basically, one assumes the following.

I-1. The zero-mean sequences $\{y(k), \boldsymbol{x}_k\}$ are related via a linear model of the form

$$y(k) = \boldsymbol{x}_k^T \boldsymbol{w}_* + v(k) \tag{3}$$

for some unknown $\boldsymbol{w}_*$, and where $v(k)$ is zero-mean with variance σ_v^2. In addition, it is assumed that the sequences $\{\boldsymbol{x}_k\}$ and $\{v(k)\}$ are independent and identically distributed, and that they are also independent of each other.

In applications, the assumption of independent regressor vectors is seldom satisfied. For example, in channel equalization, the vectors $\boldsymbol{x}_k$ are formed from a delay line (see (6)), thus $\boldsymbol{x}_n$ shares all but one of the elements of $\boldsymbol{x}_{n+1}$ so that $\{\boldsymbol{x}_n, \boldsymbol{x}_{n+1}\}$ are clearly dependent.

What makes the results obtained with the independence analysis still useful is a result first obtained in [7], showing that when the step-size is infinitesimally small ($\mu \approx 0$), the results obtained using the independence assumptions are good approximations for the actual performance of the LMS algorithm. The conclusions of [7] were later extended to more general settings (but still restricted to the LMS algorithm) in [8, 9, 10]. Similar results for other adaptive algorithms can be obtained using averaging theory and/or the ODE method (e.g., [11, 12, 13, 14]).

1.2 Performance with Faster Adaptation

The above works give a good understanding of the behavior of the LMS algorithm when the step-size μ is sufficiently small (without in fact quantifying how small μ should be). However, from the recursion (1) one can deduce that the rate of convergence of the algorithm is greatly affected by the choice of the step-size. An infinitesimal step-size ($\mu \approx 0$) implies that the weight estimates $\boldsymbol{w}_k$ change very slowly at each iteration, and consequently the convergence rate is small. This can be particularly annoying in nonstationary environments, where very slow convergence may not allow the algorithm to properly track time variations in signal statistics. A designer might then wish to employ larger step-sizes to improve the convergence speed of the algorithm, especially during the initial convergence phase (before steady-state is achieved).

The following questions are therefore relevant and remain largely unanswered.

(i) How small must the step-size be so that the independence-based approximations are still reasonable? Also, for a given value of the step-size, what is the order of magnitude of the error incurred by using these approximations ?

(ii) What is the real performance of the adaptive algorithm when the step-size is not small ?

(iii) How large the step-size can get without compromising filter (mean-square) stability?

(iv) What is the step-size that gives the fastest convergence rate ?

For step-sizes that are not infinitesimally small, there are essentially no results in the literature that predict or confirm the behavior/stability of the LMS algorithm and its variants (see, e.g., the statement in [5] regarding this issue).

The purpose of this paper is to propose a method to study the stability of the LMS algorithm, without relying on the independence assumptions and without assuming beforehand that the step-size is vanishingly small. Since an exact expression for the largest step-size (say, $\mu_{\max}$) that addresses point (iii) above is difficult to obtain, we instead derive an upper bound on how large the step-size μ can be for mean-square stability (say, $\mu < \bar{\mu}$). While the bound $\bar{\mu}$ is not tight (i.e., close to $\mu_{\max}$) at this stage of our analysis, it is, to the authors' knowledge, the first such bound and is also applicable to a generic distribution of the input sequence (and in particular, it even allows for a normally distributed input). The significance of this work is therefore in developing a framework that studies filter stability for step-sizes that are not necessarily infinitesimally small. We are currently pursuing extensions of our analysis in order to obtain tighter upper bounds and to handle larger step-sizes.

Our discussion builds upon an approach originally suggested in [15], and which will lead naturally to a state-space framework. Basically, the arguments we employ in the future sections can be summarized as follows. We first find a dynamic state-space model for the evolution of the covariance matrix $\mathrm{E}\,\tilde{\boldsymbol{w}}_k \tilde{\boldsymbol{w}}_k^T$; the states of this model will consist of the entries of the covariance matrix in addition to several other relevant quantities. The state equation will be of the form

$$\boldsymbol{\Gamma}_{k+1} = \Phi \boldsymbol{\Gamma}_k + \boldsymbol{b}, \tag{4}$$

where $\boldsymbol{b}$ is a constant vector, $\boldsymbol{\Gamma}_k$ is the state vector, and Φ is a constant matrix. With this model, the largest step-size ($\mu_{\max}$) that guarantees stable performance of the LMS filter (and therefore answers point (iii) above) will be the largest μ for which Φ is still a stable matrix, i.e.,

$$\mu_{\max} \triangleq \sup\{\mu \text{ such that } \rho(\Phi) < 1\} , \tag{5}$$

where $\rho(\Phi)$ denotes the *spectral radius* of Φ, i.e., $\rho(\Phi) = \max_i |\lambda_i(\Phi)|$.

Unfortunately, determining $\mu_{\max}$ is not a trivial task for two main reasons. First, the eigenvalues of the matrix Φ depend nonlinearly on the step size μ and, secondly, the dimension of Φ grows extremely fast with the filter length (for example, for $M = 6$ the matrix has size $28,181 \times 28,181$). It is therefore computationally infeasible to work directly with Φ; the approach is feasible only for relatively small filter lengths. For this reason, reference [15] considered only the case $M = 2$ (i.e., filter with two taps), while reference [16] used the same method for orders up to $M = 6$ coupled with a numerical procedure (viz., the power method) for the evaluation of the eigenvalues of Φ. For larger filter lengths, we need to develop an alternative procedure for the estimation of $\mu_{\max}$ that does not work directly with the matrix Φ.

The approach we propose in this paper is based on the observation that the matrix Φ, although of large dimensions, is both *sparse* and *structured*. These two properties combined can be used to derive a bound on the step-size for stable performance. [Moreover, the bound will be such that it is not a function of the maximum value that $\|\boldsymbol{x}_k\|$ can attain; the result will depend only on the distribution of the input sequence, and on average quantities.]

2 Structure of the State-Space Model

In the sequel we assume that the regressors $\{\boldsymbol{x}_k\}$ arise from a tap-delay line, say

$$\boldsymbol{x}_k = \begin{bmatrix} a(k-M+1) & a(k-M+2) & \dots & a(k) \end{bmatrix}^T . \tag{6}$$

where the input sequence $\{a(k)\}$ is assumed iid, with zero mean, and moments $\sigma_p = \mathrm{E}\, a(k)^p$, for $p \geq 1$. The assumption of iid $\{a(k)\}$ implies that the variable $a(k)$ is independent of the current weight error vector, $\tilde{\boldsymbol{w}}_k$. We also assume that $\{y(k), \boldsymbol{x}_k\}$ are related via a linear model of the form (3), with a zero-mean iid noise sequence $v(k)$ of variance σ_v^2 and that is independent of the input sequence. Using the LMS update (1) and the model (3), we find that the error equation for LMS is given by

$$\tilde{\boldsymbol{w}}_{k+1} = (I - \mu \boldsymbol{x}_k \boldsymbol{x}_k^T)\tilde{\boldsymbol{w}}_k - \mu \boldsymbol{x}_k v(k) . \tag{7}$$

We are interested in conditions under which the MSD is bounded (i.e., conditions under which $\mathrm{Tr}(\mathrm{E}\, \tilde{\boldsymbol{w}}_k \tilde{\boldsymbol{w}}_k^T)$ forms a bounded sequence). This requires that we study the stability of the matrix Φ in (4) and determine conditions under which its eigenvalues are strictly inside the unit disc. It turns out that, under the above conditions, the noise sequence $\{v(k)\}$ does not influence the stability of the recursion (4) since it does not enter Φ and only affects the driving term $\boldsymbol{b}$. For this reason, in the remainder of the paper we can assume $v(k) \equiv 0$.

2.1 Obtaining the Linear Model: An Example

We present briefly the main steps of the derivation of the state-space model (4) for $M = 2$. Similar arguments hold for larger values of M.

Recall that our purpose is to determine a recursion that describes the evolution of the entries of the covariance matrix $\mathrm{E}\,\tilde{w}_k\tilde{w}_k^T$ (whose trace determines the MSD). We denote these entries by

$$\mathrm{E}\,\tilde{w}_k\tilde{w}_k^T \triangleq \begin{bmatrix} \gamma_1(k) & \gamma_3(k) \\ \gamma_3(k) & \gamma_2(k) \end{bmatrix}.$$

The variables $\{\gamma_i(k)\}$ will be some of the elements in the state vector $\mathbf{\Gamma}_k$. This is because recursions for the $\{\gamma_i(k)\}$ will require that we also propagate other quantities. We do not show how to find all the variables in $\mathbf{\Gamma}_k$ here. This example only explains why other variables are necessary, and shows how to obtain a recursion for one of them. Similar derivations can be used for the other variables.

Let $\{\tilde{w}_{k,1}, \tilde{w}_{k,2}\}$ denote the entries of the error $\tilde{w}_k$, i.e., $\tilde{w}_k = \begin{bmatrix} \tilde{w}_{k,1} & \tilde{w}_{k,2} \end{bmatrix}^T$. Now using the error equation (7), and the the assumptions on $a(k)$, we find that

$$\gamma_1(k+1) = \gamma_1(k) - 2\mu\,\mathrm{E}\big(a(k-1)^2\tilde{w}_{k,1}^2\big) + \mu^2\,\mathrm{E}\big(a(k-1)^4\tilde{w}_{k,1}^2\big) + \mu^2\sigma_2\,\mathrm{E}\big(a(k-1)^2\tilde{w}_{k,2}^2\big).$$

The right-hand side of this recursion depends on the additional terms

$$\mathrm{E}\big(a(k-1)^2\tilde{w}_{k,1}^2\big),\quad \mathrm{E}\big(a(k-1)^4\tilde{w}_{k,1}^2\big),\quad \mathrm{E}\big(a(k-1)^2\tilde{w}_{k,2}^2\big),\quad \text{and}\quad \mathrm{E}\big(a(k-1)^2\tilde{w}_{k,1}\tilde{w}_{k,2}\big).$$

Hence, a complete recursion for $\gamma_1(k)$ requires that we also determine recursions for these quantities. If we denote the above terms in succession by $\{\gamma_4(k), \gamma_5(k), \gamma_6(k), \gamma_7(k)\}$, then, for example, the recursion that results for $\gamma_4(k+1)$ will be

$$\gamma_4(k+1) = \sigma_2\gamma_1(k) - 2\mu\sigma_2\gamma_4(k) + \sigma_2\gamma_5(k) + \mu^2\sigma_4\gamma_6(k).$$

The state vector in this case will therefore be of dimension 7, $\mathbf{\Gamma}_k = \begin{bmatrix} \gamma_1(k) & \dots & \gamma_7(k) \end{bmatrix}^T$. This procedure can in principle be repeated for any filter order M (and, in fact, as shown in [16], similar state-space models can be obtained even in situations where the $a(k)$ form a correlated sequence). However, the number of state variables will grow exponentially fast with the filter order, which makes it infeasible to pursue this line of reasoning for larger filter lengths (larger than 7 for example). Instead we focus on the structure of the coefficient matrix Φ.

2.2 Structure of Φ

As we mentioned above, a major drawback of the state-space model (4) is that its order (the size of Φ) grows exponentially fast with the filter length M. In this section, we highlight several structural properties of Φ and use them in the next section to study the stability of the LMS recursion without having to form any large matrices.

Indeed, it can be shown that Φ is highly sparse and has considerable structure. The sparseness of Φ was used in [16] to obtain approximations for its largest eigenvalue; however, the exponential growth of Φ limits the use of this technique to filters of smaller orders (say up to 6 or 7). To work with larger filter lengths, it is necessary to study the structure of Φ with more detail. It can be shown that Φ has the following properties (some are evident from the case $M = 2$ above, others only arise for larger values of M):

1. The number of nonzero elements in each row of Φ does not exceed $(M+1)^2$.
2. The number of rows of Φ is finite for finite M.

3. It is possible to select a small number of rows of Φ (no more than $16M-12$ different rows) such that *all other rows* of Φ are *permutations* of these rows. Consider the following *contrived* example,

$$\Phi = \begin{bmatrix} 1 & -2\mu & 0 & 0 & \mu^2\sigma_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \mu^2\sigma_4 & 0 & 0 & 0 & -2\mu \\ \sigma_2 & 0 & -2\mu\sigma_2 & 0 & \mu^2\sigma_6 & 0 & 0 & 0 & 0 \\ 0 & -2\mu\sigma_2 & 0 & 0 & 0 & 0 & \mu^2\sigma_6 & \sigma_2 & 0 \\ 0 & 1 & 0 & -2\mu & 0 & \mu^2\sigma_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -2\mu & \mu^2\sigma_4 & 0 \end{bmatrix}$$

The matrix Φ above has two different classes of rows. Rows numbers 1, 2, 5, and 6 form one class, and rows 3 and 4 form the other. In each class, the entries appearing in all rows are the same, although the order is different. For example, the nonzero entries in row 1 are $\mathcal{A}_1 = \{1, -2\mu, \mu^2\sigma_4\}$. The nonzero entries of rows 2, 5, and 6 also belong to $\mathcal{A}_1$. Similarly, the elements of rows 3 and 4 belong to the set $\mathcal{A}_2 = \{\sigma_2, -2\mu\sigma_2, \mu^2\sigma_6\}$. We name the sets $\mathcal{A}_1$ and $\mathcal{A}_2$ the *coefficient sets* of Φ, and we say that any two rows belonging to the same class are *p-equivalent* (i.e., they are permutations of each other).

The real matrix Φ has a similar structure; all its rows take their nonzero entries from one of (at most) $16M-12$ different coefficient sets. The coefficient sets of Φ are listed below for $M > 2$ (the integer p lies in the interval $1 \le p \le 4(M-1)$ – we can also specify the number of times each term will appear on a given row):

$$\begin{aligned}
\mathcal{A}_1 &= \Big\{1,\ -2\mu,\ \mu^2\sigma_2,\ 2\mu^2,\ \mu^2,\ -2\mu\sigma_1,\ 2\mu^2\sigma_1\Big\}, \\
\mathcal{A}_2 &= \Big\{1 - 2\mu\sigma_2 + \mu^2\sigma_4,\ \ -2\mu(\sigma_1 - \mu\sigma_3),\ \ \mu^2\sigma_2\Big\} \\
\mathcal{A}_3 &= \Big\{1,\ -\mu,\ \mu^2,\ -\mu\sigma_1,\ 2\mu^2\sigma_1,\ \mu^2\sigma_2\Big\} \\
\mathcal{A}_4 &= \Big\{1 - \mu\sigma_2,\ \ -\mu(1 - 2\mu\sigma_2),\ \ -\mu\sigma_1,\ \ \mu^2\sigma_1,\ \ \mu^2\sigma_3\Big\} \\
\mathcal{A}_{1,p} &= \Big\{\sigma_p,\ -2\mu\sigma_p,\ \mu^2\sigma_{p+2},\ 2\mu^2\sigma_p,\ \mu^2\sigma_p,\ -2\mu\sigma_{p+1},\ 2\mu^2\sigma_{p+1}\Big\} \\
\mathcal{A}_{2,p} &= \Big\{\sigma_p - 2\mu\sigma_{p+2} + \mu^2\sigma_{p+4},\ \ -2\mu(\sigma_{p+1} - \sigma_{p+3}),\ \ \mu^2\sigma_{p+2}\Big\} \\
\mathcal{A}_{3,p} &= \Big\{\sigma_p,\ -\mu\sigma_p,\ \mu^2\sigma_p,\ -\mu\sigma_{p+1},\ 2\mu^2\sigma_{p+1},\ \mu^2\sigma_{p+2}\Big\} \\
\mathcal{A}_{4,p} &= \Big\{\sigma_p - \mu\sigma_{p+2},\ \ -\mu(\sigma_p - 2\mu\sigma_{p+2}),\ \ -\mu\sigma_{p+1},\ \ \mu^2\sigma_{p+1},\ \ \mu^2\sigma_{p+3}\Big\}\ .
\end{aligned}$$

4. Those rows of Φ that have unity entries can in fact have only a single unity entry. Moreover, some of the rows of Φ have ones on the main diagonal and these are generated only by the coefficient sets $\{\mathcal{A}_1, \mathcal{A}_3\}$. For each such row with a unity entry on the main diagonal, there exists a single row from either sets $\{\mathcal{A}_{1,2}, \mathcal{A}_{3,2}\}$ (i.e., with $p=2$) with entries in the same column positions and which can be obtained from it by the following substitutions (compare the elements in the sets $\mathcal{A}_1$ and $\mathcal{A}_{1,p}$, as well as $\mathcal{A}_3$ and $\mathcal{A}_{3,p}$):

$$1 \leftarrow \sigma_2\ ,\quad \sigma_1 \leftarrow \sigma_3\ ,\quad \sigma_2 \leftarrow \sigma_4\ .$$

5. Pick a row with unity on its main diagonal and the corresponding row with the σ_2 entry in the same column position as the one on the diagonal. We can always

associate with these two rows, by means of suitable permutations, a square block whose size can vary from 2 up to M and which has the following generic form (shown here for sizes 4 and 3, and denoted by B_4 and B_3, respectively):

$$B_4 \triangleq \begin{bmatrix} 1 & -2\mu & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \sigma_2 & -2\mu\sigma_2 & 0 & 0 \end{bmatrix}, \quad B_3 \triangleq \begin{bmatrix} 1 & -2\mu & 0 \\ 0 & 0 & 1 \\ \sigma_2 & -2\mu\sigma_2 & 0 \end{bmatrix}. \tag{8}$$

The relevant characteristic of each block is the position of the 1's: one is on the main diagonal while the others are on the upper diagonal.

3 Stability of the LMS Filter

Now that several properties of Φ have been highlighted, we return to the problem of deriving a bound for the step-size that guarantees a uniformly bounded covariance matrix $\mathrm{E}\,\tilde{w}_k\tilde{w}_k^T$ (*cf.* (5)).

As mentioned before, the eigenvalues of Φ depend nonlinearly on μ and, hence, it is mathematically intractable to derive an expression for the spectral radius of Φ in terms of μ. For this reason, we shall proceed via an alternative route. For any $M \times M$ square matrix A, it holds that its spectral radius is upper bounded by its ∞-norm [18], i.e., $\rho(A) \leq \|A\|_\infty$, where $\|A\|_\infty$ is the maximum absolute row sum of A. Therefore, if we can find a μ that guarantees $\|\Phi(\mu)\|_\infty < 1$, then it also guarantees $\rho\big(\Phi(\mu)\big) < 1$.

We mentioned in the previous section that the rows of Φ belong to $16M - 12$ classes of p-equivalent rows, and that each row has only $O(M^2)$ nonzero entries. Hence, in principle, $\|\Phi\|_\infty$ can be evaluated with little effort. Unfortunately, however, several rows of Φ contain the element $\{1\}$, which makes $\|\Phi\|_\infty$ always larger than 1 regardless of μ.

To overcome this difficulty, we propose to work with a modified matrix Φ that has the same spectral radius as the original Φ but smaller ∞-norm. We achieve this by showing how to construct a similarity transformation T such that there exists a $\bar{\mu} > 0$ satisfying

$$\|T^{-1}\Phi(\mu)T\|_\infty < 1 \quad \text{for all } \mu < \bar{\mu}. \tag{9}$$

This is useful because, in view of the fact that similarity transformations preserve eigenvalues, the above $\bar{\mu}$ will also guarantee $\rho\big(\Phi(\mu)\big) < 1$ for all $\mu < \bar{\mu}$.

Special care must be taken while constructing the similarity transformation in order to preserve as much structure as possible and in order to keep the ∞-norm of $T^{-1}\Phi_M(\mu)T$ easily computable. We cannot reproduce here all the details involved in the construction procedure. We only highlight the main steps.

We construct T as a sequence of elementary similarity transformations T_k. Each one is defined once the special blocks B_k that we mentioned in (8) above have been identified. So consider a generic block B_k, which we regard as the leading block in a submatrix with k rows and multiple columns, say (for a block of size 4)

$$\left[\begin{array}{cccc|ccccccc} \boxed{1} & -2\mu & 0 & 0 & -2\mu & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & \boxed{1} & 0 & 0 & -2\mu & -2\mu & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & \boxed{1} & 0 & 0 & 0 & 0 & -2\mu & -2\mu & \ldots \\ \sigma_2 & -2\mu\sigma_2 & 0 & 0 & -2\mu\sigma_2 & 0 & 0 & 0 & 0 & 0 & \ldots \end{array}\right]$$

Recall that each block B_k has a row with a 1 on the main diagonal and $k - 2$ rows with a 1 in the upper diagonal. Our task is to replace all the 1's with smaller terms. This can be achieved by a similarity transformation T_k that is defined as follows:

1) Rows that have a 1 on the main diagonal are replaced by rows that have $1 - 2\mu\sigma_2$ on the diagonal. All other elements are $O(\mu^2)$.

2) Rows that have a 1 in off-diagonal positions are replaced by rows that have $O(\mu^{1/3})$ on these positions. All other elements will be $O(\mu^{2/3})$.

For example, the above matrix will become after the transformation T_k,

$$\left[\begin{array}{cccc|ccccccc} \boxed{1-2\mu\sigma_2} & 0 & 0 & 0 & -4\mu^2\sigma_2 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & \boxed{\mu^{1/3}} & 0 & 0 & -2\mu^{2/3} & -2\mu^{2/3} & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & \boxed{\mu^{1/3}} & 0 & 0 & 0 & 0 & -2\mu & -2\mu & \ldots \\ \mu^{-1/3}\sigma_2 & 0 & 2\mu^{2/3}\sigma_2 & 0 & & 0 & 0 & 0 & 0 & 0 & \ldots \end{array}\right]$$

Once all such transformations are carried out, the absolute row sum of each row will be strictly less than one. It then becomes possible to compute an upper bound $\bar{\mu} \leq \mu_{\max}$ by seeking the largest value of μ that results in an $\infty-$norm for the transformed matrix $\|T^{-1}\Phi(\bar{\mu})T\|_\infty$ that is less than one. This step leads to the following optimization problem (with $1 \leq p \leq 4(M-1)$),

$$\bar{\mu} \; = \; \max\,(\mu_a\mu_f) \tag{10}$$

subject to the scalars $\{\mu_f, \mu_a\}$ satisfying the inequalities (11) to (16) below:

$$\begin{aligned} &\left|1 - 2\mu_f\mu_a\sigma_2\right| + (M-1)^2\mu_f^2\left|1 - 2\mu_f\mu_a\sigma_2\right| + \mu_f^2\left|\mu_a\sigma_2 - 2\mu_f\mu_a^2\sigma_4\right| + \\ &+2(M-2)(M-1)^2\mu_f^2 + 2(M-2)(M+1)\mu_f^2\mu_a\sigma_2 + (M-2)(M-1)^3\mu_f^3 + \\ &+5(M-2)(M-1)\mu_f^3\mu_a\sigma_2 + 2(M-2)\mu_f^3\mu_a^{3/2}\sigma_3 < 1, \end{aligned} \tag{11}$$

$$\begin{aligned} &\left|1 - 2\mu_f\mu_a\sigma_2\right| + 2(M-1)(M^2-3M+1)\mu_f^2 + \\ &+2(M^2-M-3)\mu_f^2\mu_a\sigma_2 + (M-1)^2(M^2-3M+1)\mu_f^3 + \\ &+(5M^2-15M+9)\mu_f^3\mu_a\sigma_2 + 2(M-2)\mu_f^3\mu_a^{3/2}\sigma_3 < 1. \end{aligned} \tag{12}$$

$$\mu_f^{\frac{1}{3(M-2)}} + 2(M-1)\mu_f^{\frac{2}{3}} + (M-1)^2\mu_f^{\frac{5}{6}} + \mu_f^{\frac{5}{6}}\mu_a\sigma_2 < 1, \tag{13}$$

$$\left|1 - 2\mu_f\mu_a\sigma_2 + \mu_f^2\mu_a^2\sigma_4\right| + (M-1)^2\mu_f^2\mu_a\sigma_2 + 2(M-1)\mu_f^2\mu_a^{3/2}\sigma_3 < 1, \tag{14}$$

$$\begin{aligned} &\mu_a^{p/2}\sigma_p + (M^2-3M+2)\mu_f\mu_a^{p/2}\sigma_p + (M-1)^2\mu_f^2\mu_a^{p/2}\sigma_p + \\ &+2\mu_f\mu_a^{(p+1)/2}\sigma_{p+1} + 2(M-1)\mu_f^2\mu_a^{(p+1)/2}\sigma_{p+1} + \mu_f^2\mu_a^{(p+2)/2}\sigma_{p+2} < \mu_f^{\frac{1}{3}}, \end{aligned} \tag{15}$$

$$\begin{aligned} &\left|\mu_a^{p/2}\sigma_p - 2\mu_f\mu_a^{(p+2)/2}\sigma_{p+2} + \mu_f^2\mu_a^{(p+4)/2}\sigma_{p+4}\right| + (M-1)^2\mu_f^2\mu_a^{(p+2)/2}\sigma_{p+2} + \\ &+2(M-1)\mu_f^2\mu_a^{(p+3)/2}\sigma_{p+3} < 1 \end{aligned} \tag{16}$$

The solid curve in Figure 1 shows a plot of the solution $\bar{\mu}$ of (10) for a sequence $\{a(k)\}$ that is normally distributed with $\sigma_2 = 0.01$. The rate of decay of $\bar{\mu}$ is dependent on the signal distribution and is proportional to $1/M^5$ in this example, as indicated by the broken line.

While a tighter bound for $\bar{\mu}$ that decays slower with M can be obtained by careful selection of more specialized similarity transformations T_k, the purpose of this work has been to suggest a framework for the derivation of such bounds and for the stability analysis of adaptive schemes without prior assumption of vanishingly small step-sizes.

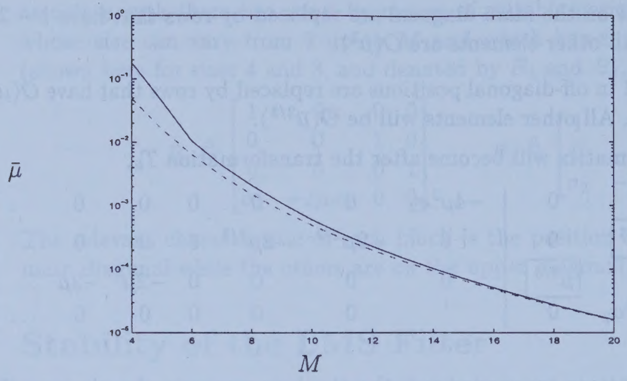


Figure 1: A plot of $\bar{\mu}$ versus filter size M for a Gaussian distribution.

4 Some Concluding Remarks

There are still several issues that remain to be addressed. These include weakening the assumptions on $\{a(k)\}$ (such as allowing for correlated input sequence), obtaining a more relaxed bound on the step-size (one that decreases slower with M), and weakening the condition on the noise $\{v(k)\}$. These issues are currently under investigation.

Another issue of interest and which is discussed in the companion article [17] is the following. Once an analysis is performed on the expected behavior of an adaptive scheme (as we did above), it is common in practice to confirm the results by means of simulations by generating so-called ensemble-average learning curves. These curves are usually obtained by averaging over no more than 100-200 repeated experiments and they provide approximations for the evolution of the error measures $\mathrm{E}\, e(k)^2$ or $\mathrm{E}\, \|\tilde{w}_k\|^2$ as a function of time. It has been observed in practice that these averaged curves tend to match theoretical results reasonably well for sufficiently small step-sizes and that, therefore, they tend to provide a good approximation for the expected performance of an adaptive filter.

For larger step-sizes, however, we observed that care must be taken in trying to validate or predict the performance of an adaptive scheme by means of ensemble-average learning curves. It can be shown that, for larger μ, the number of simulations that must be performed in order to obtain a good approximation to $\mathrm{E}\, e(k)^2$ or $\mathrm{E}\, \|\tilde{w}_k\|^2$ can be of the order of tens or hundreds of thousands, depending on the distribution of the input sequence. A smaller number of simulations can lead to erroneous or deceptive conclusions. An analytical explanation and more details on this effect are given in [17].

References

[1] B. Widrow et al. Stationary and nonstationary learning characteristics of the LMS adaptive filter. *Proceedings of the IEEE*, 64:1151–1162, 1976.

[2] B. Widrow and S. D. Stearns. Adaptive Signal Processing. Prentice-Hall, NJ, 1985.

[3] S. Haykin. Adaptive Filter Theory. Prentice-Hall, 3rd edition, NJ, 1996.

[4] A. Feuer and E. Weinstein. Convergence analysis of LMS filters with uncorrelated Gaussian data. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33(1):222–229, February 1985.

[5] D. T. M. Slock. On the convergence behavior of the LMS and the normalized LMS algorithms. *IEEE Transactions on Signal Processing*, 41(9):2811–2825, September 1993.

[6] K. Mayyas and T. Aboulnasr. The leaky LMS algorithm: MSE analysis for gaussian data. *IEEE Transactions on Signal Processing*, 45(4):927–934, April 1997.

[7] J. E. Mazo. On the independence theory of equalizer convergence. *The Bell System Technical Journal*, 58:963–993, May–June 1979.

[8] S. K. Jones, R. K. Cavin, and W. M. Reed. Analysis of error-gradient adaptive linear estimators for a class of stationary dependent processes. *IEEE Transactions on Information Theory*, IT-28:318–329, March 1982.

[9] O. Macchi and E. Eweda. Second-order convergence analysis of stochastic adaptive linear filtering. *IEEE Transactions on Automatic Control*, AC-28(1):76–85, January 1983.

[10] O. Macchi. Adaptive Processing: The LMS Approach with Applications in Transmission. Wiley, NY, 1995.

[11] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22:551–575, August 1977.

[12] W. A. Sethares, B. D. O. Anderson, and C. R. Johnson. Adaptive algorithms with filtered regressor and filtered error. *Mathematics of Control, Signals, and Systems*, 2:381–403, 1989.

[13] V. Solo and X. Kong. *Adaptive Signal Processing Algorithms.* Prentice-Hall, Englewood Cliffs, NJ, 1995.

[14] H. J. Kushner and F. J. Vázquez-Abad. Stochastic approximation methods for systems over an infinite horizon. *SIAM Journal of Control and Optimization*, 34(2):712–756, March 1996.

[15] S. Florian and A. Feuer. Performance analysis of the LMS algorithm with a tapped delay line (two-dimensional case). *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(6):1542–1549, December 1986.

[16] S. C. Douglas and W. Pan. Exact expectation analysis of the LMS adaptive filter. *IEEE Transactions on Signal Processing*, 43(12):2863–2871, December 1995.

[17] V. H. Nascimento and A. H. Sayed. Are ensemble-average learning curves reliable in evaluating the performance of adaptive filters? In *Proc. Asilomar Conference on Signals, Systems, and Computers*, CA, 1998.

[18] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1987.

Blind Adaptive Cross-Pole Interference Cancellation Using Fractionally-Spaced CMA

Wonzoo Chung[*], John Treichler[†], and C. Richard Johnson, Jr.[*]

[*]wonzoo(johnson)@ee.cornell.edu [†] jrt@appsig.com

School of Elec. Eng.
Cornell University
Ithaca, NY 14853

Applied Signal Technology
Sunnyvale, CA 94086

Abstract

Cross-pole interference (CPI) cancellation is a specific type of source separation problem in a high-rate digital communication system which utilizes dual polarized carriers in order to achieve frequency reuse. In this paper, we analyze a fractionally-spaced (FS) blind adaptive cross-pole interference (CPI) cancellation scheme using the constant modulus algorithm (CMA). First, we study zero forcing solutions and minimum mean-squared error solutions of FS CPI equalizer on a cross coupled channel with a certain diversity. Second, we show close analogies between the FS-CMA CPI equalizer and the conventional FS-CMA blind adaptive equalizer. Thus existing results on FS-CMA can be applied to describe the important behavior of the FS-CMA CPI equalizer, for example, the agreement between CM-solutions and zero-forcing (ZF) solutions in the absence of channel noise and perfect modelling and the robustness of the FS-CMA CPI equalizer in more realistic situations (wchich include channel noise and undermodelling). Finally, we address the initialization of this blind algorithm by using a novel expression of the FS-CMA CPI cost function.

1 Introduction

In modern communication systems, the demand for transmitting extensive information through channels of limited bandwidth has encouraged the search for an efficient transmission scheme exploiting both spectral and spatial resources. Especially in digital microwave radio systems, such an efficient system can be realized by frequency reuse of two orthogonally polarized carrier waves at the same frequency. Two mutally independent source sequences $\mathbf{s}_1$ and $\mathbf{s}_2$ are transmitted simultaneously on vertical and horizontal polarizations at the same nominal carrier frequency and received by two appropriately polarized antennas. Armed with this technique, theoretically, one can transmit twice the amount of information through a fixed bandwidth channel. However, it has been observed that in practice the performance of the system tends to be degraded due to the interference (including cross-pole) of signals passing through physical channels.

---

[*]Supported by NSF Grants MIP-9509011 and ECS-9528363 and Applied Signal Technology.

Interference is caused by not only the bandwidth of the channel (inter symbol interference (ISI)) but also by the structure of orthogonal polarization: refraction and reflection of carrier waves during propagation allow only mixed signals to reach the receiver antennas. The structural interference is referred to as *cross-polarization interference* (CPI) and modelled as a linear addition of two source sequences. Recovering signals despite CPI, as well as ISI, is essentially a source separation problem. One may apply conventional dual finite impulse response (FIR) filter equalization techniques based on estimation of second order statistics estimation or training sequence adaptation [7, 8].

A blind adaptive CPI cancellation scheme based on high order source statistics has been proposed by Treichler et al. and implemented recently [10, 11, 12]. This scheme is a natural extension of the constant modulus algorithm (CMA) to a dual FIR filter equalizer. It also can be applied to other blind signal separation problems in digital communication systems, such as sidelobe canceling. Although in experiments CMA CPI shows satisfactory robust performance [12], one issue yet to be resolved is the initialization of this blind algorithm. In this paper, we consider CPI cancellation with a fractionally spaced equalizer and present an analysis of fractionally spaced CMA CPI focusing on the effect of initialization.

In Section 2, we establish a fractionally spaced CPI cancellation model. With a generalized version of a conventional perfect equalization assumption on a fractionally spaced receiver, FS CPI with an FIR channel model admits FIR equalizer solutions. We study zero-forcing (ZF) solutions and minimum mean squared error (MMSE) solutions in Section 3. FS CPI equalization using CMA is analyzed by analogy to conventional FS CMA equalization and the initialization issue is addressed in Section 4.

2 Fractionally Spaced CPI Cancellation

2.1 System model

Figure 1 models cross-polarization interference in a communication system using vertically and horizontally polarized antennas. Two independent source sequences $\mathbf{s}_1$ and $\mathbf{s}_2$ are transmitted on horizontally and vertically polarized carrier waves, respectively. FIR filters $\mathbf{c}_{11}$ and $\mathbf{c}_{22}$ represent the direct path between horizontal and vertical polarization antennas, respectively, while $\mathbf{c}_{12}$ and $\mathbf{c}_{21}$ represent cross-coupling terms.

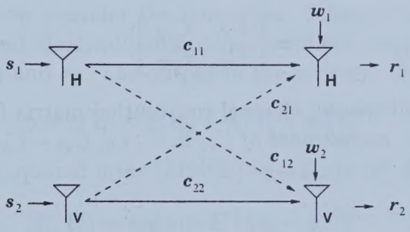


Figure 1. Cross Polarization Interference

The received signal of a fractionally spaced receiver is sampled at a fraction of the baud rate. This way, additional channel information from the fractional sampling can be utilized for equalization. We consider a $T/2$-spaced receiver (i.e., the sampling interval is half of the baud interval T). A $T/2$-spaced CPI equalizer intended to extract one of the transmitted signals is illustrated in Figure 2.

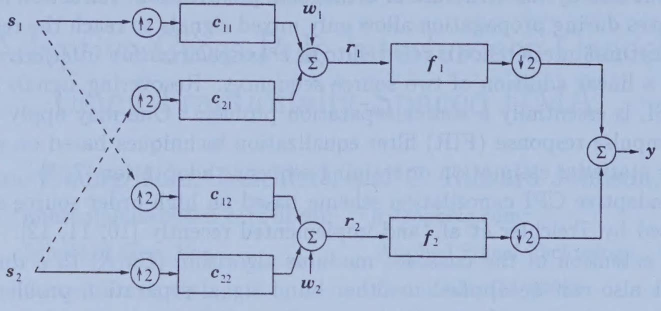


Figure 2. FS CPI equalizer for a desired signal.

For now suppose there is no channel noise $\mathbf{w}_1$ or $\mathbf{w}_2$. Then, for channel paths $\mathbf{c}_{ij}$ and equalizers $\mathbf{f}_k$, the combined baud spaced response can be written as a linear combination of channel convolution matrices C_{ij} and equalizer vectors $\mathbf{f}_i$, i.e., $C_{ij}\mathbf{f}_k$ [5]. Thus in the absence of noise, the output y of the model in Figure 2 is given by

$$y = \mathbf{s}_1^T(C_{11}\mathbf{f}_1 + C_{12}\mathbf{f}_2) + \mathbf{s}_2^T(C_{21}\mathbf{f}_1 + C_{22}\mathbf{f}_2) \tag{1}$$

where C_{11}, C_{12}, C_{21}, and C_{22} are the (row-decimated, baud spaced) convolutional matrices of $\mathbf{c}_{11}, \mathbf{c}_{12}, \mathbf{c}_{21}$ and $\mathbf{c}_{22}$ respectively. We denote the system response vectors of each channel-equalizer combination as

$$\begin{aligned} \mathbf{h}_1 &:= C_{11}\mathbf{f}_1, & \mathbf{h}_2 &:= C_{21}\mathbf{f}_1 \\ \mathbf{g}_1 &:= C_{12}\mathbf{f}_2, & \mathbf{g}_2 &:= C_{22}\mathbf{f}_2 \end{aligned} \tag{2}$$

We define $\mathbf{q}_1 := \mathbf{h}_1 + \mathbf{g}_1$ and $\mathbf{q}_2 := \mathbf{h}_2 + \mathbf{g}_2$ as the system responses of the FS CPI equalizer. for sources $\mathbf{s}_1$ and $\mathbf{s}_2$, and we have the following linear relation:

$$\begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \end{bmatrix} := \begin{bmatrix} \mathbf{h}_1 + \mathbf{g}_1 \\ \mathbf{h}_2 + \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{bmatrix} \tag{3}$$

Define

$$C := \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \tag{4}$$

as an analogy of the (baud spaced) channel convolution matrix for an FS equalizer. C is invertible when the *Schur complement of* C_{11} in C, i.e, $C_{22} - C_{21}C_{11}^{-1}C_{12}$, is non singular [6].

2.2 Assumptions

With a fractionally-spaced equalizer, it has been shown that perfect equalization is possible [9]. We assume the following [5] for the channels $\{\mathbf{c}_{ij}\}$ and the equalizers $\{\mathbf{f}_i\}$ so that C can be invertible (perfect equalization):

A1. The $T/2$-channel models $\{\mathbf{c}_{ij}\}$ are all the same length and satisfy the sub-channel disparity condition by having no roots common to among its sub-channels.

A2. The length of $\{\mathbf{f}_i\}$ satisfies the length condition of perfect equalization for $T/2$-spaced equalization, i.e. the length of equalizer $N_{f_i} = N_c - 2$.

From these assumptions we have invertible channel convolution matrices C_{11}, C_{12}, C_{21}, and C_{22} [5]. Now $\mathbf{h}_2$ and $\mathbf{g}_2$ can be expressed in terms of $\mathbf{h}_1$ and $\mathbf{g}_1$ respectively:

$$\mathbf{h}_2 = A\mathbf{h}_1, \quad \mathbf{g}_2 = B\mathbf{g}_1, \tag{5}$$

where $A := C_{21}C_{11}^{-1}$ and $B := C_{22}C_{12}^{-1}$. The non-singularity of the Schur complement of C_{11}, i.e. $C_{22} - C_{21}C_{11}^{-1}C_{12}$, is equivalent to the non-singularity of $A - B$. We assume that our cross-pole channels have diversity such that

A3. $A - B$ is a non-singular matrix, which guarantees that C is invertible.

3 Second-Order-Statistics Solutions

3.1 Zero-Forcing Solutions

In the absence of noise, the output of the CPI equalizer is given by

$$y = \mathbf{s}_1^T(\mathbf{h}_1 + \mathbf{g}_1) + \mathbf{s}_2^T(\mathbf{h}_2 + \mathbf{g}_2) \tag{6}$$

To equate y to (a delayed version of) $\mathbf{s}_1$, we have two equations

$$\left\{ \begin{array}{l} \mathbf{h}_1 + \mathbf{g}_1 = \mathbf{e}_i \\ \mathbf{h}_2 + \mathbf{g}_2 = 0 \end{array} \right. \quad \text{or equivalently,} \quad \begin{bmatrix} \mathbf{e}_i \\ 0 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{bmatrix} \tag{7}$$

Since we assume that C is invertible, (7) yields a unique solution depending on the chosen delay $\mathbf{e}_i := [0 \cdots 0 \; 1(i-th) \; 0 \cdots 0]^T$.

$$\left\{ \begin{array}{l} \mathbf{f}_1^* = C_{11}^{-1}(B - A)^{-1}B\mathbf{e}_i \\ \mathbf{f}_2^* = -C_{12}^{-1}B^{-1}AC_{11}\mathbf{f}_1^* \end{array} \right. \tag{8}$$

The derivation of $\mathbf{f}_i^*$ can be found in [6].

3.2 MMSE solution

In this subsection, we now consider the noisy case. Suppose the noises w_1 and w_2 are mutually independent and identically distributed with the same variance σ_w^2, and uncorrelated to the sources s_1 and s_2. The output in the presence of noise is

$$y = \mathbf{s}_1^T\mathbf{q}_1 + \mathbf{s}_2^T\mathbf{q}_2 + \mathbf{w}_1^T\mathbf{f}_1 + \mathbf{w}_2^T\mathbf{f}_2 \tag{9}$$

The minimum mean squared error (MMSE) equalizers $(\mathbf{f}_1^*, \mathbf{f}_2^*)$ are defined as

$$(\mathbf{f}_1^*, \mathbf{f}_2^*) = \arg\min_{\mathbf{f}_1, \mathbf{f}_2} E\left\{\|s_1 - y\|^2\right\} \tag{10}$$

Let $\mathbf{f} := [\mathbf{f}_1^T \; \mathbf{f}_2^T]^T = [f_{1_0}, \cdots, f_{1_{N-1}}, f_{2_0}, \cdots f_{2_{N-1}}]^T$. Since $\mathbf{s}_1$ and $\mathbf{s}_2$ are mutally independent, we treat $\mathbf{s} := [\mathbf{s}_1^T \; \mathbf{s}_2^T]^T$ as a i.i.d. source, and the same for the noise $\mathbf{w} := [\mathbf{w}_1^T \; \mathbf{w}_2^T]^T$. Now we have simplified the expression of y which is similar to that for the FS MMSE equalizer [5].

$$y = \mathbf{s}^T C\mathbf{f} + \mathbf{w}^T\mathbf{f} \tag{11}$$

The MMSE solution $\mathbf{f}^*$ is given by [5]

$$\mathbf{f}^* = (C^H C + \lambda I)^{-1} C^H \mathbf{e}_i \tag{12}$$

where $\lambda = \sigma_w^2/\sigma_s^2$. Or, we can obtain a more explicit expression in terms of C_{ij} by considering:

$$MSE = E\left\{\|s_1 - y\|^2\right\} = \sigma_s^2(\|\mathbf{h}_1 + \mathbf{g}_1 - \mathbf{e}_i\|^2 + \|A\mathbf{h}_1 + B\mathbf{g}_1\|^2) + \sigma_w^2(\|\mathbf{f}_1\| + \|\mathbf{f}_2\|) \tag{13}$$

By setting $\frac{\partial}{\partial h_1} MSE$ and $\frac{\partial}{\partial g_1} MSE$ to zero, we obtain the following equations.

$$\left\{\begin{array}{l} (I + A^H A + \lambda C_{11}^{-H} C_{11}^{-1})\mathbf{h}_1^* + (I + A^H B)\mathbf{g}_1^* - \mathbf{e}_i = 0 \\ (I + B^H B + \lambda C_{12}^{-H} C_{22}^{-1})\mathbf{g}_1^* + (I + B^H A)\mathbf{h}_1^* - \mathbf{e}_i = 0 \end{array}\right. \tag{14}$$

Because of the non symmetric terms $A^T B$ and $B^T A$, the solutions cannot be expressed in a simple form. From some rather straightforward calculations, we obtain:

$$\begin{aligned} \mathbf{f}_1^* &= C_{11}^{-1} \frac{I - (I + A^H A + \lambda C_{11}^{-H} C_{11})^{-1}}{(I + A^H A + \lambda C_{11}^{-H} C_{11})^{-1} - (I + B^H B + \lambda C_{12}^{-H} C_{12}^{-1})^{-1}(I + A^H B)} \mathbf{e}_i & (15) \\ \mathbf{f}_2^* &= C_{12}^{-1} \frac{A^H A - B^H A + \lambda C_{11}^{-H} C_{11}^{-1}}{B^H B - A^H B + \lambda C_{12}^{-H} C_{12}^{-1}} C_{11} \mathbf{f}_1^* & (16) \end{aligned}$$

where the fraction denotes left inversion (i.e. $\frac{P}{Q} := Q^{-1}P$). Note that the MMSE solutions in (15) and (16) agree with ZF solutions in (8) in the absence of noise (i.e. $\lambda = 0$).

If a communication system allows training signals, the MMSE solutions shown above can be obtained via the LMS algorithm:

$$\begin{aligned} \mathbf{f}_1^{(n+1)} &= \mathbf{f}_1^{(n)} - \mu(y - s_1)\mathbf{r}_1^* & (17) \\ \mathbf{f}_2^{(n+1)} &= \mathbf{f}_2^{(n)} - \mu(y - s_1)\mathbf{r}_2^* & (18) \end{aligned}$$

where μ is a step size and $\mathbf{r}_i^*$ is the complex conjugated regressor vector of equalizer $\mathbf{f}_i$. Notice that the desired signal is determined by the training sequence, i.e. $\mathbf{s}_1$ could be replaced by $\mathbf{s}_2$.

4 FS-CM Blind Cross-Polarized Interference Equalizer

In many applications a blind equalization scheme is preferred, since training signals degrade channel capacity or may not be available. However, blind equalization based on second-order statistics tends to lack robustness in realistic situations [3]. Adaptive blind equalization using the Constant Modulus Algorithm has gained great interest in industry due to its robustness.

The application of CMA to CPI is a natural extension of CMA which inherits the robustness of CMA. To introduce the CMA CPI equalizer, we review CMA in the following subsection.

4.1 Constant Modulus Algorithm

The constant modulus algorithm is a blind adaptive algorithm which exploits the non-gaussian nature of a typical digital communication source [5]. CMA is designed to minimize the following cost function

$$J(\mathbf{f}) = \frac{1}{4} E \left\{ (|y(n)|^2 - \gamma)^2 \right\} \tag{19}$$

with the stochastic gradient update equation

$$\mathbf{f}^{(n+1)} = \mathbf{f}^{(n)} - \mu \frac{\partial}{\partial \mathbf{f}} J(\mathbf{f})|_{\mathbf{f}=\mathbf{f}^{(n)}} = \mathbf{f}^{(n)} - \mu y(|y(n)|^2 - \gamma)\mathbf{r}^* \tag{20}$$

where μ is a step size and $\mathbf{r}^*$ denotes the complex conjugated regressor of the equalizer $\mathbf{f}$.

In a fractionally-spaced system, the cost function can be expressed in terms of the baud-spaced system response $\mathbf{h}$ (assuming a BPSK source without loss of generality)[1]:

$$4J_H(\mathbf{h}) = (||\mathbf{h}||^2 - 1)^2 + \eta_s I_{\mathbf{h}} \tag{21}$$

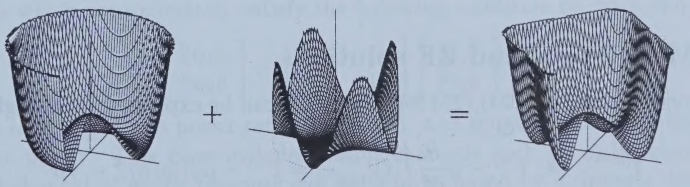
where $I_{\mathbf{h}}$ is a semi-definite real-valued function measuring inter-symbol interference induced by the system response $\mathbf{h}$ defined as

$$I_{\mathbf{h}} := \sum_{i \neq j}^{N-1} |h_i|^2 |h_j|^2 \tag{22}$$

and η_s characterizes the deviation of the source kurtosis κ_s from that κ_w of a Gaussian sequence.

$$\eta_s := (\kappa_w - \kappa_s) \tag{23}$$

Expression (21) can be understood as a radial and spherical decomposition of the CMA cost function.



a) Radial $(||\mathbf{h}||^2 - 1)^2$ b) Spherical $\eta_s I_{\mathbf{h}}$ c) FS-CMA $(||\mathbf{h}||^2 - 1)^2 + \eta_s I_{\mathbf{h}}$

Figure 3. Spherical and radial decomposition of CMA cost function

From this representation, some robustness features of CMA can be shown [1, 2, 4]:

- Bad local minima (with large noise gain) tend to disappear as noise increases
- Regions of convergence of good local minima (with small noise gain) expand near the origin, while those of bad local minima shrink.

4.2 CMA CPI equalizer

The structure of the CM CPI equalizer from [12] is shown in Figure 4.

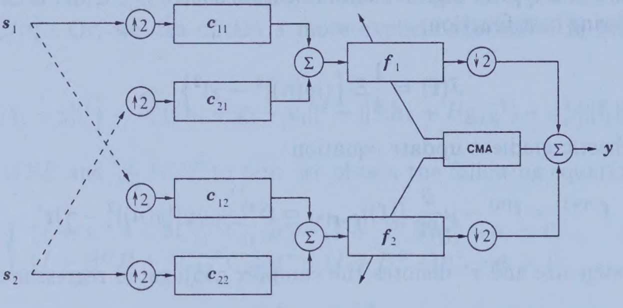


Figure 4. FS CPI cancellation using CMA.

The constant modulus algorithm for CPI uses two stochastic gradient updates: one each for $\mathbf{f}_1$ and $\mathbf{f}_2$.

$$\mathbf{f}_1^{(n+1)} = \mathbf{f}_1^{(n)} - \mu \frac{\partial}{\partial \mathbf{f}_1} J(\mathbf{f}_1, \mathbf{f}_2)|_{\mathbf{f}_1=\mathbf{f}_1^{(n)}, \mathbf{f}_2=\mathbf{f}_2^{(n)}} = \mathbf{f}_1^{(n)} - \mu y(|y(n)|^2 - \gamma)\mathbf{r}_1^* \tag{24}$$

$$\mathbf{f}_2^{(n+1)} = \mathbf{f}_2^{(n)} - \mu \frac{\partial}{\partial \mathbf{f}_2} J(\mathbf{f}_1, \mathbf{f}_2)|_{\mathbf{f}_1=\mathbf{f}_1^{(n)}, \mathbf{f}_2=\mathbf{f}_2^{(n)}} = \mathbf{f}_2^{(n)} - \mu y(|y(n)|^2 - \gamma)\mathbf{r}_2^* \tag{25}$$

where the cost function J is the same as in (19), and $\mathbf{r}_1$ and $\mathbf{r}_2$ denote the regressors of $\mathbf{f}_1$ and $\mathbf{f}_2$ respectively.

The blind nature of CMA raises the following important questions:

- Where are the CM local minima located ?
- How should the equalizer be initialized in order to get a (particular) desired output ?

In the following subsections we answer the above questions.

4.3 CM solutions and ZF solutions

The two update equations (24),(25) for CMA CPI can be expressed by a single equation:

$$\mathbf{f}^{(n+1)} = \mathbf{f}^{(n)} - \mu \frac{\partial}{\partial \mathbf{f}} J(\mathbf{f})|_{\mathbf{f}=\mathbf{f}_1^{(n)}} = \mathbf{f}^{(n)} - \mu y(|y(n)|^2 - \gamma)\mathbf{r}^* \tag{26}$$

where $\mathbf{r} = [\mathbf{r}_1^T\ \mathbf{r}_2^T]^T$ and $\mathbf{f} = [\mathbf{f}_1^T\ \mathbf{f}_2^T]^T$. Given (3), $\mathbf{q}^T := [\mathbf{q}_1^T\ \mathbf{q}_2^T]^T$. Then $\mathbf{q} = C\mathbf{f}$ and we have

$$y = \mathbf{s}^T\mathbf{q} \tag{27}$$

where $\mathbf{s} = [\mathbf{s}_1^T\ \mathbf{s}_2^T]^T$. This expression of the output results in a conventional CMA cost function (assuming a BPSK source) of the form

$$J(\mathbf{q}) = (||\mathbf{q}||^2 - 1)^2 + \eta_s I_\mathbf{q} \tag{28}$$

with a corresponding update equation. Depending on the initialization of the equalizer $\mathbf{f}_1$ and $\mathbf{f}_2$, we have $\mathbf{q}_1 = \mathbf{e}_i$ and $\mathbf{q}_2 = 0$ or vice versa. In the absence of noise, these CMA solutions match the ZF solutions.

The region of convergence for each delay δ_i is characterized by the matrix $C = [C_{11}C_{12}; C_{21}C_{22}]$ [2]. Since we are interested in the initialization not for each delay but for a desired signal, we consider another expression of the CMA CPI cost function:

FS-CMA CPI Cost Function: The FS-CMA CPI cost function can be expressed in terms of $\mathbf{q}_1$ and $\mathbf{q}_2$:

$$J = \left(\|\mathbf{q}_1\|^2 + \|\mathbf{q}_2\|^2 - 1\right)^2 + 2\eta_s\|\mathbf{q}_1\|^2\|\mathbf{q}_2\|^2 + \eta_s I_{\mathbf{q}_1} + \eta_s I_{\mathbf{q}_2} \tag{29}$$

(29) can be interpreted as follows:

1) Ignoring $\eta_s I_{\mathbf{q}_1} + \eta_s I_{\mathbf{q}_2}$ terms, J is a CM cost function of $\|\mathbf{q}_1\|$ and $\|\mathbf{q}_2\|$. It implies

$$\begin{array}{l} \|\mathbf{q}_1\| = 1 \text{ and } \|\mathbf{q}_2\| = 0 \text{ or} \\ \|\mathbf{q}_1\| = 0 \text{ and } \|\mathbf{q}_2\| = 1 \end{array} \tag{30}$$

2) For $\mathbf{q}_2 = 0$ ($\mathbf{q}_1 = 0$), J is a conventional CMA cost function for $\mathbf{q}_1$ ($\mathbf{q}_2$). Thus we have $\mathbf{q}_1 = \mathbf{e}_i$ and $\mathbf{q}_2 = 0$ or vice versa.

This interpretation indicates that the norm of $\mathbf{q}_1$ and $\mathbf{q}_2$ would be of primary concern in the initialization of CPI CMA, which reduces a multi-dimensional problem into a two dimensional problem.

4.4 Initialization Issue

To describe a closed form expression for the regions of convergence of a given desired signal is almost impossible because of the complicated fourth order nature of the cost function. In this paper we limit our analysis to a special case that matches the experimental setup in [12]. From the observation that CMA CPI can be roughly understood as a CMA on $\|\mathbf{q}_1\|$ and $\|\mathbf{q}_2\|$, we establish regions of convergence for a desired signal in terms of $\|\mathbf{f}_1\|$ and $\|\mathbf{f}_2\|$. To do that, we need a strong assumption on the channel. We consider a class of channels which approximately satisfy the following condition on $\|\mathbf{q}_1\|$ and $\|\mathbf{q}_2\|$.

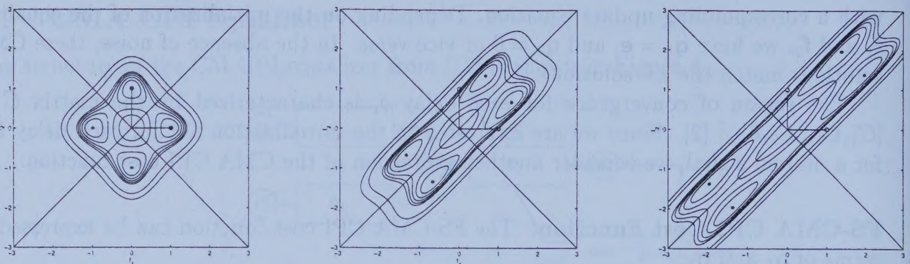
$$\begin{bmatrix} \|\mathbf{q}_1\| \\ \|\mathbf{q}_2\| \end{bmatrix} = \begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix} \begin{bmatrix} \|\mathbf{f}_1\| \\ \|\mathbf{f}_2\| \end{bmatrix} \tag{31}$$

where α is labeled as cross power ratio, $0 \leq \alpha \leq 1$, and roughly represents the cross-pole interference power. This class includes simple channels such as scalar channels and a wide band channel of which response can be approximated by a discrete delta function with delays. Let's denote $\bar{C} := [1\ -\alpha\ ; -\alpha\ 1]$.

The regions of convergence of CMA for $\|\mathbf{q}_1\|$ and $\|\mathbf{q}_2\|$ are equally divided hyper-cones (Figure 5 a) and those for $\|\mathbf{f}_1\|$ and $\|\mathbf{f}_2\|$ are characterized by singular decomposition of $\bar{C}^{-1}$ ("rotation" and "dilation") [2]

$$\begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{bmatrix} \begin{bmatrix} \frac{1}{1+\alpha} & 0 \\ 0 & \frac{1}{1-\alpha} \end{bmatrix} \begin{bmatrix} \cos(\frac{\pi}{4}) & \sin(\frac{\pi}{4}) \\ -\sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{bmatrix}$$

The cost function is stretched by $\frac{1}{1-\alpha}$ in $x+y$ direction and shrunk by $\frac{1}{1+\alpha}$ in $x-y$ direction, and thus the noise gain of equalizer ($\|f_k\|$) increases. However, the symmetry of $\bar{C}^{-1}$ keeps the regions of convergence unchanged (Figure 5).



a) $\alpha = 0$ b) $\alpha = 0.5$ c) $\alpha = 0.7$

Figure 5. Regions of convergence for $||\mathbf{f}_1||$ and $||\mathbf{f}_2||$ (consider only for $||\mathbf{f}_i|| \geq 0$).

This result implies a equalizer setting $||\mathbf{f}_1|| = 1$ and $||\mathbf{f}_2|| = 0$ is still valid (in it's convergence to s_1) for channels satisfying above assumptions. Furthermore, from the behavior of $\frac{1}{1-\alpha}$ as $\alpha \rightarrow 1$, the FS-CMA CPI equalizer under these channel assumption shows remarkable robustness of bit error rate (BER) performace when the cross-pole diversity (the condition number of matrix $A - B$ in Asumption 3) is degraded: Notice that in Figure 6 the BER is uniformly under 10^{-5} until about 10dB cross power ratio.

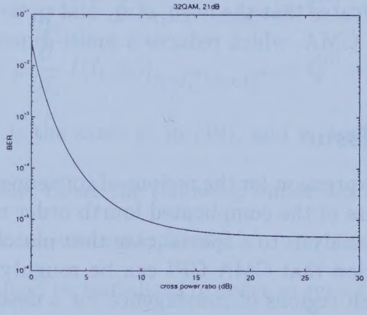


Figure 6. BER and cross power ratio α (32 QAM at 21 dB noise) from (31)

Since our analysis was performed under the assumption of a noise-free channel, the above plot is not precise. However, the experimental result in [12] shows exactly the same trend of Figure 6, because the effect of mild noise on the CMA cost function can be understood as "smoothing" the cost surface.

5 Conclusion

We have studied CPI equalizers of a factionally spaced system focusing on the blind FS-CMA CPI equalizer. We have shown that the solutions of the FS-CMA CPI equalizer match the ZF-solutions in the absence of noise. The FS-CMA CPI equalizer inherits the robust properties of conventional CMA equalizer (e.g. with noise and undermodelling). BER performance and initialization sensitivity have been shown to be robust to lack of corss-pole diversity for a simple channel class. Study of initialization for more general channels have yet to be studied.

References

[1] W. Chung, J. P. LeBlanc, "The local minima of fractionally-spaced CMA blind equalizer cost function in the presence of channel noise," *Proc. ICASSP*, May 1998.

[2] W. Chung, C.R. Johnson, "Characterization of the regions of convergence of CMA adaptive blind fractionally spaced equalizer," *to appear in Asilomar 1998*

[3] T.J. Endres, B.D.O. Anderson, C.R. Johnson, Jr. and L. Tong, "On the robustness of FIR channel identification from fractionally-spaced received signal second-order-statistics," *IEEE Signal Processing Letters,* vol. 3, no. 5., pp. 153-155, May 1996.

[4] I. Fijalkow, A. Touzni, and J. R. Treicher, "Fractionally spaced equalization using CMA: robustness to channel noise and lack of disparity," *IEEE Transactions on Signal Processing,* vol. 45, no.1 pp. 55-66, Jan. 1997.

[5] C. R. Johnson, Jr., P. Schniter, T. J. Endres, J. Behm, D. R. Brown, and R. A. Casas, "Blind equalization using the constant modulus criterion: A review" *Proc. IEEE,* Oct. 1998

[6] P. Lancaster and M. Tismenetsky, *The Theory of Matrices,* Academic Press, 1985

[7] J. Namiki and S. Takahara, "Adaptive receiver for cross-polarized digital transmission," in *Proc. ICC* 1981, June 1982, pp.46.31-46.35.

[8] J. Salz, "Data transmission over cross-coupled linear channels," *AT&T Technical Journal,* vol. 64, No. 6. July-August 1985, pp. 1147-1160.

[9] L. Tong, G. Xu, T. Kailath, "Blind identification and equalization based on second-order statistics," *IEEE Trans. on Info. Theory,* vol. IT-40, pp. 340-349, Mar. 1994

[10] J.R. Treichler and M.G. Larimore, "New processing techniques based on the constant modulus adaptive alogrithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-33, No. 2, April 1985, pp. 420-431.

[11] J.R. Treichler and J. Bohanon, "Blind demodulation of high-order QAM signals in the presence of cross-pole interference," *Proceedings of the 1998 MILCOM Conference,* Monterey, CA, June 3, 1998

[12] J.R. Treichler, J. Bohanon, M. Larimore, G. Collins, V. Wolff, C.R. Johnson, and S.L. Wood, "Blind separation and combination of high-rate digital QAM radio signals," *Proc. 8th IEEE Digital Signal Processing Workshop*, Bryce, UT, August 1998.

An Adaptive Solution to the *El Farol* Problem

William A Sethares
Dept. of Electrical and Computer Engineering
University of Wisconsin
Madison WI 53706-1691
sethares@ece.wisc.edu

Ann M. Bell
Dept. of Economics
Vanderbilt University
Nasville TN 37212
ann.m.bell@vanderbilt.edu

Abstract

The *El Farol* problem provides a simple paradigm for congestion and coordination problems such as arise with over utilization of the Internet. This paper reviews the *El Farol* problem and surveys previous solutions, which typically involve complex learning algorithms. A simple adaptive strategy is proposed, and the strategy is investigated via simulation. The algorithm is analyzed in a few simple cases.

1 The *El Farol* Problem

El Farol is a bar in Santa Fe. The bar is popular, but becomes overcrowded when more than sixty people attend on any given evening. Everyone enjoys themselves when fewer than sixty people go, but no one has a good time when the bar is overcrowded. How can (or how do) people choose whether to go to *El Farol* on any given evening?

If there was a centralized social planner with the power to dictate who could go (and who would not be allowed to go), then the planner could choose exactly sixty people each night. The bar would never be undercrowded (insuring that a many people as possible enjoyed themselves) and it would never be overcrowded (so no one would ever have a bad time). In this sense, "social welfare" would be maximized or optimized.

But *El Farol* has no social planner, and everyone must make their own decision whether to go or not. Brian Arthur [1] observed that the number of people at the bar tends to hover near sixty. How is it possible for so many independent decision makers, each working with incomplete information, to arrive at an aggregate solution that is so near to optimal? This is the *El Farol* "problem".

El Farol is a simplification of a large class of congestion and coordination problems that arise in modern engineering and economic systems. For instance, consider logging onto an internet provider such as America On Line. If only a few people are also online, then emails and downloads occur rapidly and the user has a "good time". But if too many people are online, then service is slow, graphics files are tedious to download, and the user has a bad experience. When and whether to log on again will be determined, at least in part, by recent experiences.

El Farol is a model of congestion because more people want to attend on any given evening than the bar can happily accommodate. In this sense, it is like the bottleneck problems investigated by Vickrey [8]. Vickrey found that by instituting a graduated toll at the bottleneck (for instance, at a bridge with a restricted number of lanes), overall social welfare could be increased because those with the willingness to pay the toll would do so, leaving others to utilize the bottleneck at non-peak hours. This is one way of smoothing the demand for a scarce good, and the success of this method in some (small number of road) facilities has encouraged some recent researchers (see [7] and references therein) to propose that similar fees be charged to internet users.

But *El Farol* is also a model of coordination (and coordination failure when too many people arrive). In this sense it is a kind of extended "prisoners dilemma" [6]. Each agent wants to go to the bar every evening, but if everyone goes, then all have a bad time. Only by coordinating their efforts can everyone be made better off.

2 Solutions to *El Farol*

Arthur [1] first proposed *El Farol* as an example of a problem that required boundedly rational decision making [6]. His aim was to mimic the kinds of decision making that people use in complex and uncertain environments, and he endowed his agents with the ability to exercise a simple kind of inductive reasoning.

In Arthur's solution, agents do their best to predict how many others will attend *El Farol* each time. If their prediction is less than sixty then they go, while if their prediction is greater than sixty then they stay at home. There are many methods of making such a prediction: simple averages, moving averages, linear or nonlinear filters, various "rules of thumb", etc. Each of Arthur's one hundred agents make a variety of different predictions, and then "believe" the prediction that was correct most frequently in the recent past. When he simulated his bar-going society, he found that indeed, the population at the bar tended to hover near 60.

Figure 1, taken from [1], shows a sample of the kinds of behavior Arthur observed in his model. First, the mean attendance is very close to 60. About half of the time there are more than 60 people at the bar (having a bad time). The variance of the attendance is large, often being greater than 70 or less than 50. Looking closely at the behavior of individual agents, Arthur says, "... the predictors self-organize into an equilibrium pattern or 'ecology' in which the active predictors... average 40% above 60, 60% below 60. This emergent ecology is almost organic in nature."

At the heart of the *El Farol* problem, according to Arthur, is that many of the agents must be wrong in their predictions, no matter how the predictions are made. To see this, suppose that 61 agents choose to attend the bar on a given evening; these 61 incorrectly predicted that fewer than 60 would attend. On the other hand, suppose that 59 attended. Then the other 41 incorrectly predicted that more than 60 would attend. Thus on average, roughly half the agents are wrong every time. These repeated incorrect predictions drive the continual churning of rules that Arthur observes in his simulations; agents switch between different methods of predicting because no single rule works in all cases. Indeed, suppose all agents adopted the same 'successful' strategy. Then they would all make the same predictions and would all go (or not go) together – ensuring that no one enjoys the experience. Thus Arthur argues that agents must utilize a heterogeneous set of rules.

Edmonds [3] expands on Arthur's approach by endowing the agents with an expanded set of predictors, and by allowing communication between the agents (including the ability for agents to 'lie' to each other). The resulting simulations are similar to Arthurs, in that agents continually change their prediction strategies and that an average of about 60 people attend each evening.

Casti [2] uses Arthurs formulation of *El Farol* as a springboard for more general speculation: to frame the question of defining a complex adaptive system with intelligent agents using only local information. Casti states "The upshot of the *El Farol* problem is that there exists no deductive chain of reasoning that will enable the individual to best decide whether to go to the bar or not."

In this paper, we show that it is not necessary for agents to utilize different rules, nor for agents to constantly switch between rules when deciding whether to attend *El Farol.* We do this by demonstrating a simple adaptive scheme, which, if followed by all the agents results in 'convergence' in the sense that roughly 60 attend each evening. The next section motivates this adaptive strategy and details the algorithm. Section 4 examines the generic behavior of the adaptive strategy, as demonstrated by simulations. This is made more concrete by the analysis of section 5, though it is only possible to prove the behavior of the algorithm in certain simple cases. The final section places the adaptive solution in the broader context of Arthurs original intent (that of a simple paradigm for bounded rationality) and suggests several areas for further investigation.

3 An Adaptive Solution

It is tautological that people prefer to experience good times rather than bad, to repeat the enjoyable and to minimize the unpleasant. Though the *El Farol* situation provides a simple setting in which good and bad are clearly defined, it is not possible to know in advance whether a trip to the bar will be good or bad, since this depends on the actions of everyone else.

Consider the situation from the point of view of an agent who must decide whether to attend or not. (Note that, unlike Arthur, we do not require agents to explicitly predict how many others will go.) Suppose that the agent is initially willing to risk going once every c times. This is reasonable since agents prefer to go rather than to stay at home, but prefer staying at home to a crowded bar. There are two possibilities when the agent goes to the bar. Consistent with the desire to maximize pleasure and minimize painful experiences, the agent is willing to go more often (to decrease c) if the bar is uncrowded, but prefers to go less often if the bar is crowded. Over time, the agent gathers information about the state of the bar, and 'remembers' this in the form of the parameter c.

Observe the similarity between this simple adaptive strategy for the *El Farol* problem and the "number guessing game" of the first chapter of Johnson [5]. In both cases, there is an unknown. If new information reveals that the current guess is too low (or too high), then the guess is increased (or decreased). The simplicity of the scheme suggests that it may be feasible to analyze the resulting behavior, though the current situation is more complicated since the "true" value of the unknown depends directly on everyone's guesses.

To make the transition from uncrowded to crowded less drastic, a 'don't care' or 'dead zone' parameter d can be incorporated. The bar will be considered crowded if more than 60 agents attend, uncrowded if $60 - d$ or fewer attend, and moderate if between. In response, the agents will modify their c value only when outside the moderate region. This better models the deterioration in service due to congestion. For instance, the download time of a file does not necessarily change dramatically under the load of a single new customer, but can deteriorate rapidly when a group of new customers suddenly demand service.

To write this strategy in symbolic form, suppose there are M agents competing for the N possible spaces at *El Farol*, and let d represent the size of the dead zone. (In Arthur's simulations, $M = 100$ and $N = 60$, but he does not explicitely state how his agents handle the transition from crowded to uncrowded.) Let c_i be the frequency (period) with which the ith agent wishes to attend, let p_i, the 'phase', be the number of timesteps until the ith agent attends next (thus p_i is always less than c_i), and let μ_i be a characteristic parameter that defines how much the ith agent changes c_i in response to new information. Let k be a time (iteration) counter. Since the c_i and p_i change as time evolves, let $c_i(k)$ and $p_i(k)$ designate the instantaneous values of c_i and p_i at the time k. Thus the pair $[c_i(k), p_i(k)]$ represents the state of the ith agent at time k, and the concatenation of all M pairs gives the state vector for the entire system.

Let

$$N(k) = \sum_{i=1}^{M} 1_{[p_i(k) \leq 0]} \tag{1}$$

count the number of agents attending at time k, where 1_A is the indicator function that is one if the event A is true and zero otherwise. The evolution of the $c_i(k)$ and $p_i(k)$ is defined by

$$c_i(k+1) = \begin{bmatrix} c_i(k) & \text{if } p_i(k) > 0 \\ \text{Max}(\mu_i,\ c_i(k) + \mu_i \ \text{Dsgn}(N(k) - N)) & \text{if } p_i(k) \leq 0 \end{bmatrix}$$
$$p_i(k+1) = \begin{bmatrix} p_i(k) - 1 & \text{if } p_i(k) > 0 \\ c_i(k) & \text{if } p_i(k) \leq 0 \end{bmatrix}$$

where $\text{Dsgn}(x)$ represents the signum function with dead zone that is positive when $x > 0$, negative when $x - d \leq 0$, and zero otherwise.

The operation of this algorithm is uncomplicated. The 'phase' term p_i begins at c_i and counts down until it reaches zero. Once zero, it is reset to the current value of c_i. This is repeated over and over. Meanwhile, the counter c_i remains unchanged, except when p_i hits zero, which represents the time when the agent attends the bar. At this point, c_i is increased by μ_i if the bar is crowded and decreased by μ_i if the bar is uncrowded. The $\text{Max}(\cdot)$ function insures that the counter always remains positive. The algorithm can be written more succinctly as

$$\begin{aligned} c_i(k+1) &= \text{Max}(\mu_i,\ c_i(k) + \mu_i \ 1_{[p_i(k) \leq 0]} \ \text{Dsgn}(N(k) - N)) \\ p_i(k+1) &= (p_i(k) - 1) \ 1_{[p_i(k) > 0]} \ + \ c_i(k) \ 1_{[p_i(k) \leq 0]} \end{aligned} \tag{2}$$

This algorithm can also be 'derived' as an approximation to an instantaneous gradient descent for the minimization of the cost function

$$J(k) = |N(k) - N|. \tag{3}$$

The typical gradient strategy is to update the state by

$$c_i(k+1) = c_i(k) - \mu_i \frac{dJ(k)}{dc_i(k)}.$$

Ignoring the singularity at $N(k) = N$, this leads to

$$c_i(k+1) = c_i(k) - \mu_i \ \text{sgn}(N(k) - N) \ \frac{dN(k)}{dc_i(k)}. \tag{4}$$

Unfortunately, $\frac{dN(k)}{dc_i(k)}$ is not well defined. To proceed, observe that larger $c_i(k)$ imply that $1_{[p_i(k) \leq 0]}$ is one less frequently, and hence that $N(k)$ tends to be smaller. Similarly, smaller $c_i(k)$ imply that $1_{[p_i(k) \leq 0]}$ is one more frequently, and hence that $N(k)$ tends to be larger. In other words, the sign of this 'derivative' is negative. Thus (4) can be rewritten

$$c_i(k+1) = c_i(k) + \mu_i \ \text{sgn}(N(k) - N)$$

which holds whenever $1_{[p_i(k)\leq 0]}$. Combining this with the phase gives the complete algorithm (2). The 'sgn' function can be readily generalized to incorporate the dead zone by suitable modification of the cost function (3).

Alternatively, one could consider the squared cost function $J(k) = (N(k) - N)^2$, which leads to the LMS-like [4] update

$$c_i(k+1) = c_i(k) + \mu_i \ (N(k) - N).$$

The full algorithm would then be identical to (2) with the Dsgn(·) function removed. By analogy, (2) is a kind of "double signed" LMS; one sign function is explicit, the other arises because only the sign of the derivative in (4) is known.

4 Generic Behavior of the Adaptive Solution

This section explores the generic behavior of the system when each of the $M = 100$ agents follows the strategy defined by (2) above. Though details of the various simulations differ, a typical case is illustrated in Figure 2. The counters $p_i(0)$ were initialized randomly and the distribution of the $p_i(0)$'s was chosen so that too few people attend the bar initially. Initializing so that too many agents attend at the start exhibits the same kind of behavior.

Perhaps the most striking aspect of these simulations is that the number of attendees converges very rapidly (within a few hundred iterations) to near the optimal value of $N = 60$. Once the initial transient has passed, there is only an occasional spike above 60. The population hovers very near the optimal that could be achieved by a central planner, despite the fact that each agent is autonomous, and makes the decision to go (or not to go) based on local information, that is, on its own experiences. In comparison, in Arthur's setup, there are far greater excursions about the optimal value and the bar is overcrowded about half the time.

Perhaps even more surprising than the number of agents going to the bar is the makeup of the crowd. Figure 3 shows values of the counters $c_i(K)$ at the final iteration $K = 2000$ of a typical simulation run. (As suggested by the apparent 'asymptotic' performance, the initial conditions are irrelevant to such histograms.) The agents have divided themselves into two groups. About 55 (in this case 54) agents have $c_i(K)$ less than 1, indicating that they go to *El Farol* every time. The remaining agents have much larger $c_i(K)$, and vie with each other for the remaining open slots. The occasional spikes (when the number of attendees is above 60) in Figure 2 represents times when several of the remaining agents happen to go simultaneously.

Thus the population has segregated itself into a group of 'regulars' who go every time and a group of 'casuals' who only attend occasionally. Both groups are satisfied. The regulars can truthfully say that they go to *El Farol* all the time and that it is almost never crowded. The casuals can truthfully say that they go sometimes, but that when they try to go more frequently, it always seems crowded. This division of the population appears nowhere in the algorithm statement; rather, it is an emergent property of the adaptive solution to the *El Farol* problem. When agents follow this adaptive strategy, *El Farol* looks more like *Cheers*.

The final simulation, shown in Figure 4, investigates the effect of allowing the agents to update their counters at every iteration, whether they have personally attended the bar or not. (This more closely reflects the information structure in Arthur's simulations.) Observe that the variance about $N = 60$ is much larger than in Figure 2, and there is a cloud of values around 30 indicating that seats in the bar often remain unfilled.

Somewhat paradoxically, the adaptive solution thus functions better when agents have restricted information.

5 Analysis of the Adaptive Solution

Though it may be possible to ultimately understand the complete behavior of the adaptive solution to the *El Farol* problem, this paper has more modest goals. We characterize the equilibra of the system, and then demonstrate the behavior of the algorithm in a simple setting which shows that, under certain circumstances, the casuals may attend less and less frequently, their counters diverging towards infinity, albeit at an ever slower rate.

In a strict sense, there can be no equilibria for the system defined by (2) with $M > N$ because the phase terms p_i continually decrease from c_i to zero, only to be reinitialized at c_i. However, from the point of view of the population at the bar (the only externally observable quantity), if the c_i have reached equilibrium, then $N(k)$ is completely predictable. We formalize this as

Proposition 5.1 *Suppose that* $c_i(k) = c_i^*$ *for all* $k \geq K$. *Then* $N(k)$ *is periodic for* $k \geq K$.
Proof: Since the $c_i(k)$ *are fixed,* $1_{[p_i(k) \leq 0]}$ *is periodic with period* c_i^*. *Hence the sum of such indicator functions is periodic with period (at most)* $c^* = \prod_{i=1}^{M} c_i^*$, *and* $N(k)$ *is* c^* *periodic.*

In the following, when the c_i portion of the state has converged, we say that the system (2) has reached equilibrium, with the understanding that this does not imply a constant attendance at the bar, but rather a periodic pattern of attendance, in accordance with the proposition.

To get an idea of the number of possible equilibria, consider a single 'slot' of the M possible. This could be occupied by a customer who attends every time. It could also be occupied by two customers who alternate, i.e., who have different phases. In general, any number could alternate on successive evenings, and still only occupy a single space at the bar.

For example, consider the $M = 5$, $N = 3$, $d = 1$ case. (These three parameters determine the number of possible equilibria.) Equilibria of (2) are related to the number of different ways that M can be partitioned into sums of integers. For instance, $M = 5$ can be partitioned in six ways

$$\begin{array}{llllll} (a) & 1+1+1+1+1 & (b) & 2+1+1+1 & (c) & 2+2+1 \\ (d) & 3+1+1 & (e) & 3+2 & (f) & 4+1 \end{array}$$

Let each of the summands represent (the integer part of) the c_i^* value for the agents in a candidate equilibrium configuration. For instance, a "3" means that three agents alternate filling one bar seat, while a "1" means that a single agent attends every time. If there are more than N terms in a given partition (such as in (a) and (b) above) then more than N agents attend the bar. When (2) is iterated, the Dsgn$(\cdot)$ term is positive, and the c_i will change. Hence this cannot represent an equilibrium configuration. Similarly, if there are fewer than N terms (such as in (e) and (f)), the Dsgn$(\cdot)$ term will be negative and again this is not an equilibrium. If however, the number of terms in the partition equals N (as in (c) and (d)) then all slots are exactly filled each time, Dsgn$(\cdot)$ is identically zero, and it is an equilibrium. The same argument holds in general.

To count the number of possible equlibria, we require two functions. For $M > N$, let $\phi(M, N)$ represent the number of ways that the integer M can be partitioned into a sum of exactly N integers. $\phi(5,3)$ is illustrated above. Let $\psi(n)$ represent the number of ways that 1 can be written as a sum of exactly n fractions of the form $\frac{1}{i_1}+\frac{1}{i_2}+...+\frac{1}{i_n}$ for integers i_k. $\psi(n)$ counts the number of ways that n agents can alternate to occupy a single bar stool. For instance, the "3" in (d) above can represent three agents each of whom comes every third time $\frac{1}{3}+\frac{1}{3}+\frac{1}{3}$, or it can represent a single agent who comes every second time in collaboration with two agents who come every fourth time $\frac{1}{2}+\frac{1}{4}+\frac{1}{4}$. The total number of equilibria is thus a combination of ϕ and ψ.

Proposition 5.2 *The algorithm (2) with $d=1$ has at most $\sum_{j=1}^{\phi(M,N)}\prod_{k=1}^{N}\psi(i_k^j)$ equilibria.*
Proof: As shown above, there are $\phi(M,N)$ different configurations in which the equilibria may occur. For $j=1,2,...,\phi(M,N)$, each of these has the form $i_1^j+i_2^j+...+i_N^j$. Since $\psi(n)$ counts the number of ways that n agents can alternate to occupy a single bar stool, $\prod_{k=1}^{N}\psi(i_k^j)$ counts the number of possible equilibria within each configuration. Summing over all the configurations gives the desired result.

When $d=0$, the bar is always crowded or uncrowded, Dsgn$(\cdot)$ is always sign definite, and the c_i is changed at each iteration[1]. Hence there can be no equilibria.

When $d>1$, then there are more equilibria than $\phi(M,N)$. For example, consider the case $M=5$, $N=4$, with $d=2$. Then (b)-(f) are equilibria by the same argument as above. But there are also others. The pair of agents representing the "2" in case (b) might actually consist of two agents attending every nth time (though out of phase with each other). The population of the bar would then fluctuate between three (when neither attends) and four when one attends. In neither case is there a change in the corresponding c_i values because of the nontrivial dead zone. Thus the kinds of arguments used in the previous proposition can only provide a lower bound on the number of kinds of equilibria. For general d, the number of equilibria of (2) is a nontrivial counting problem.

Given the simulations, it may appear that the system (2) must always converge to some equilibrium. We now demonstrate a simple case in which the c_i of certain agents can actually diverge.

Proposition 5.3 *Suppose M is given, $N=M-1$, $d=0$, and that $\mu_i \ll 1$ for all i. Suppose further that the system (2) has evolved (or is initialized) so that there are $M-1$ 'regulars' with $c_i<1-\mu_i$. The remaining agent, designated agent m, has $c_m \gg 1$. Then $c_m(k)\to\infty$ as $k\to\infty$.*
Proof: Suppose the mth agent first attends the bar at time k_1. The regulars attend by assumption and so it is overcrowded. Hence

$$c_i(k_1+1)=c_i(k_1)+\mu_i \text{ for all } i.$$

Because $c_i(k_1)<1-\mu_i$, $c_i(k_1+1)<1$ for all the regulars but $c_m(k_1+1)>1$. Hence only the $M-1$ regulars attend at timestep k_1+1. Thus the bar is uncrowded at timestep k_1+1 and $c_i(k_1+2)=c_i(k_1+1)-\mu_i=c_i(k_1)$ for all $i\neq m$, returning the regulars to their starting place, but leaving the mth agent with larger c_m. This exact sequence of events happens each time the mth agent attends, and hence c_m increases by μ_m each time. Since c_m directly specifies how often the mth agent attends, c_m diverges at an ever decreasing rate.

---

[1] The only exception is for 'regulars' with small μ_i. However, Since $M>N$, all agents cannot be regulars.

Such agents can increase their counters forever. Presumably, real humans in such a situation would eventually withdraw and go to some other bar. This result is actually quite singular in the sense that it is driven by a peculiar initial condition. Moreover, the assumption that $d = 0$ is crucial. If $d \geq 1$, then this mechanism for divergence fails because the casual agents can eventually synchronize their periods into simple integer multiples in which the pattern of attendence remains within the moderate region.

6 Discussions and Conclusions

While Arthur [1] first introduced *El Farol* as an example of a problem requiring boundedly rational agents, we have shown that this model is also interesting as a simple model of congestion and coordination behaviors such as occur in an overcrowded Internet.

Arthur believed that any solution to the *El Farol* problem would require heterogeneous agents, that is, agents who pursue different strategies. In contrast, we have presented a simple adaptive solution which can be followed by all agents, and which can readily solve the problem. Each agent in the adaptive solution is characterized by a parameter that determines how often the agent attends, and a stepsize that determines how much to change the parameter in response to each visit to the bar. Our agents, like Arthurs, are boundedly rational, and are completely deterministic.

The adaptive solution to the *El Farol* problem differs from Arthur's strategy in several ways. We do not require agents to make explicit predictions of the state of the bar, and we allow them to use only the information that they have readily available, i.e., their own experiences. This makes our model more realistic (it is not clear how the agents in Arthur's model learn what happened at the bar when they are absent). Finally, we allow the 'don't care' zones. Arthur is not explicit about how his agents handle the transition from uncrowded to crowded.

Arthur's solution, in which each agent maintains a bank of strategies leads to patterns of attendance that fluctuate considerably above and below the optimal. When crowded, none of the agents enjoys themselves. When undercrowded, there is a wasted resource represented by the empty seats at the bar. The adaptive solution, in contrast, leads to patterns of attendance with much smaller variance, and hence much less waste. Generically, we believe the adaptive method will lead to convergence to an optimal solution, one at which the bar is neither under nor over crowded.

The adaptive solution thus provides a simple mechanism whereby a large collection of decentralized decision makers, each acting in their own best interests and with only limited knowledge, can solve a complex congestion and social coordination problem. Moreover, convergence to the solution is relatively rapid (depending on the initial conditions) and robust. We have also demonstrated that under certain singular initial conditions it is possible for certain of the agents to 'withdraw' from the society. Nonetheless, the population of the bar remains near its optimal value.

Do we believe that customers of *El Farol* tick off the time till they can go again, increasing or decreasing a counter with each new visit? Of course not. But the incentives are in agreement with the commonsense idea that people tend to minimize bad experiences and maximize good ones. Moreover, the global behavior of the population is consistent with certain kinds of coordination phenomena. For instance, users of an Internet provider can spread demand over much of the day even though everyone might prefer (all else being equal) to log on in the middle of the afternoon. By developing certain habits (for instance, always logging on at the same time) users send signals to others to avoid these times. In this way, demand is smoothed.

There are many ways to generalize the adaptive solution to the *El Farol* problem. For instance, different people have different tolerances for what constitutes a crowd or an unacceptable delay. Each agent could also have a parameter that represents their tolerance for congestion. Additionally, to more closely model the Internet situation, one might incorporate time-of-day or day-of-week as a parameter in the process of logging on. It would also be instructive to create a hybrid situation in which a number of Arthur-like agents and a number of adaptive agents compete for spaces at the bar. We believe that the more goal directed adaptive agents will again become regulars at the bar, relegating Arthur's more haphazard agents to squabbling over the few remaining seats.

References

[1] W. B. Arthur, "Inductive reasoning and bounded rationality: the *El Farol* problem", *American Economic Review: American Economic Assocation Papers and Proceedings 1994*, Vol. 84, Pp. 406-411, May 1994.

[2] J. L. Casti, "Seeing the light at *El Farol*," *Complexity,* Vol. 1, No. 5, Pp. 7-10, 1996.

[3] B. Edmunds, "Gossip, sexual recombination and the *El Farol* bar: modelling the emergence of heterogeneity," *Proceedings of the 1998 Conference on Computation in Economics, Finance, and Engineering*, Cambridge, England, June 1998.

[4] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, NJ, 1991.

[5] C. R. Johnson Jr., *Lectures on Adaptive Parameter Estimation*, Prentice-Hall, NJ 1988.

[6] D. M. Kreps, *A Course in Microeconomic Theory,* Princeton University Press, 1990.

[7] J. K. MacKie-Mason and H. R. Varian, "Pricing the Internet", in *Public Access to the Internet*, Ed. B. Kahin and J. Keller, Prentice-Hall, Englewood Cliffs, New Jersey, 1994.

[8] W. S. Vickrey, "Congestion theory and transport investment", *American Economic Review*, No. 59, Pp. 251-260, 1969.

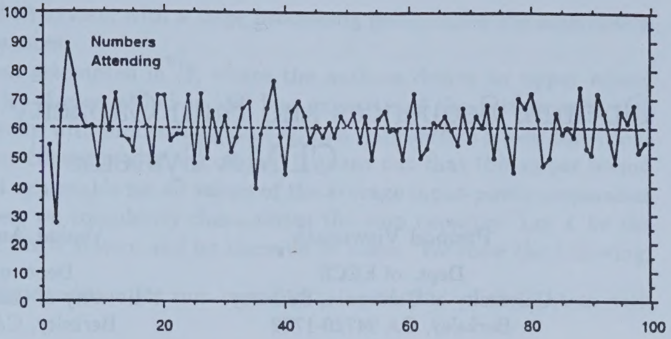


Figure 1: In Arthur's simulations, agents use a collection of predictors to decide whether to attend *El Farol*. Observe the large variance about the optimal 60 attendees.

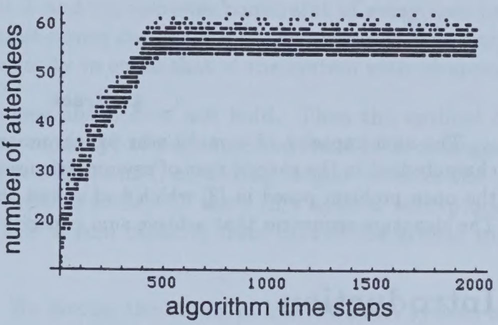


Figure 2: When all agents use the adaptive solution, the number of attendees only rarely exceeds the critical N=60.

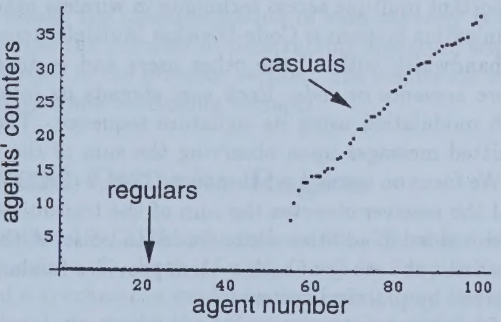


Figure 3: An emergent property of the adaptive solution is that the population divides itself into `regulars' and `casuals'.

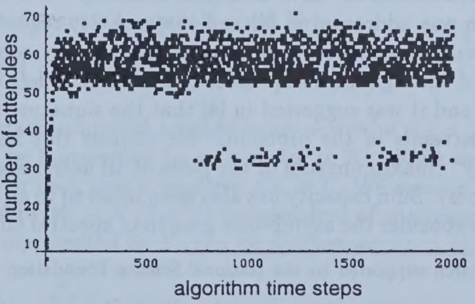


Figure 4: In this simulation, each agent updates its counter at every time step. Using more information does not necessarily lead to better results.

Optimal Sequences and Sum Capacity of Synchronous CDMA Systems *

Pramod Viswanath
Dept. of EECS
University of California, Berkeley
Berkeley, CA 94720-1772
pvi@eecs.berkeley.edu

Venkat Anantharam
Dept. of EECS
University of California, Berkeley
Berkeley, CA94720-1772
ananth@eecs.berkeley.edu

Abstract

The sum capacity of a multi-user synchronous CDMA system is completely characterized in the general case of asymmetric user power constraints - this solves the open problem posed in [3] which had solved the equal power constraint case. The signature sequences that achieve sum capacity are also identified.

1 Introduction

An important multiple access technique in wireless networks and other common channel communication systems is Code-Division Multiple Access (CDMA). Each user shares the entire bandwidth with all the other users and is distinguished from the others by its *signature sequence* or *code*. Each user spreads its information on the common channel through modulation using its signature sequence. Then, the receiver demodulates the transmitted messages upon observing the sum of the transmitted signals embedded in noise. We focus on symbol synchronous CDMA (S-CDMA) systems where in each symbol interval the receiver observes the sum of the transmitted signals in that symbol interval alone embedded in additive white Gaussian noise. Although complete synchronization is too idealistic, the study of such system provides fundamental limits on the best that can be achieved by practical systems.

Of fundamental interest in this system is the capacity region defined as the set of information rates at which users can transmit while retaining reliable transmission. This problem was addressed in [4] and the capacity region was characterized as a function of the signature sequences and average input-power constraints of the users. However, the choice of the signature sequences of the users is left open to the designer of the CDMA system and it was suggested in [4] that the signature sequences could be optimized given the constraints of the problem. We address this issue and focus on finding the "sum capacity" (maximum sum of the rates of all users; maximum over all choices of signature sequences). Sum capacity has also been referred to as "spectral efficiency" in [5] where the authors consider the asymptotic growth of spectral efficiency in a large system (with large

---

*Research supported by the National Science Foundation under grants NCR 94-22513 and IRI 95-05561.

number of users in a CDMA system with a large processing gain) under the assumption of random spreading sequences.

This problem has been attempted in [3] where the authors derive an upper bound on the sum capacity. This upper bound is just the capacity of the system with "no spreading", i.e., of the system with processing gain 1 and an appropriate power constraint on its input (sum of power constraints of the users). It turns out that this upper bound on the sum capacity is not achievable for all values of the average input-power constraints of the users. In this paper, we completely characterize the sum capacity. Let L be the processing gain of the S-CDMA system and let there be M users. We show the following:

1. $M \geq L$ is a necessary condition for sum capacity to equal that of the system with no spreading.
2. When $M \geq L$, the product of L and input-power constraint of every user being a lower bound to the sum of input-power constraints of the users is a necessary and sufficient condition for sum capacity to equal that of the system with no spreading.
3. Suppose the condition mentioned above does not hold. Then the optimal choice (in the sense of achieving sum capacity) of signature sequence is to orthogonalize the user with the loosest power constraint (i.e., having largest value of the power constraint) and the problem reduces to $M-1$ users in a system with processing gain $L-1$. There is a strict loss in sum capacity from that of the system with no spreading in this case.

This paper is organized as follows: We discuss the model of the S-CDMA system briefly and recall the characterization of the capacity region for fixed choice of signature sequences in Section 2. Our main result, the characterization of sum capacity is in Section 3. In Section 4 we briefly discuss the problem of constructing signature sequences that achieve sum capacity characterized in the previous section. The results are summarized in Section 5 which also contains some concluding remarks.

2 S-CDMA Model and Capacity Region

We consider the discrete-time, baseband S-CDMA channel model. There are M users in the system and the processing gain is L. Both M and L will be fixed throughout this paper. Since we have assumed a synchronous model we can restrict our attention to one symbol interval. As is traditional, we model the information transmitted (symbol) by each user as independent random variables $X_1, \ldots, X_M$. We assume that there is an average input-power constraint on the transmit symbols given by

$$E\left[X_i^2\right] \leq p_i \quad \forall i = 1 \ldots M$$

Let $D = \operatorname{diag}\{p_1, \ldots, p_M\}$ and the maximum total average input-power be $p_{tot} = \sum_{i=1}^{M} p_i$. Let the signature sequence of user i be represented by s_i, a vector in $\mathcal{R}^L$. Each signature sequence has power equal to L i.e., for each user i, we have $s_i^t s_i = L$. We assume an ambient white Gaussian noise, denoted by $W \sim \mathcal{N}(0, \sigma^2 I)$ independent of the transmitted symbols. Then the received signal, represented by Y, can be written as:

$$Y = \sum_{i=1}^{M} s_i X_i + W$$

Let us represent the $L \times M$ matrix $[s_1 \ldots s_M]$ by S. The S-CDMA channel above is a special case of the M-user Gaussian multiple access channel and the capacity region (set of rates at which reliable communication is possible) is well known (see Section 7 of [1]) as the closure of the convex hull of the union over all product probability densities $p_{\bar{X}}$ on the inputs $X_1, \ldots, X_M$ of the rate regions

$$\mathcal{C}(S, p_{\bar{X}}) = \bigcap_{J \subseteq \{1,\ldots,M\}} \left\{ (R_1, \ldots, R_M) : 0 \leq \sum_{i \in J} R_i \leq I(Y; X_i, i \in J | X_i, i \in J^c) \right\} \tag{1}$$

Continuing as in [3], the union region over the product distributions, as a function of S, can be simply written as

$$\mathcal{C}(S) = \bigcap_{J \subseteq \{1,\ldots,M\}} \left\{ (R_1, \ldots, R_M) : 0 \leq \sum_{i \in J} R_i \leq \frac{1}{2L} \log \left[\det \left(I + \frac{1}{\sigma^2} S_J D_J S_J^t \right) \right] \right\} \tag{2}$$

where J is a non-null set and R_i is the rate in bits/chip of user i. Here, S_J is $L \times |J|$ matrix $[s_i : i \in J]$ and D_J is the $|J| \times |J|$ matrix diag $\{p_i : i \in J\}$. Observe that the choice of product distribution for each user i, X_i is distributed as $\mathcal{N}(0, p_i)$ makes the region in (1) equal to that in (2). The sum capacity represents the maximum sum of rates of all users at which users can transmit reliably. Following the notation in [3], the sum capacity is defined formally as

$$\mathcal{C}_{\text{sum}} = \max_{S \in \mathcal{S}} \max_{R \in \mathcal{C}(S)} \sum_{i=1}^{M} R_i \tag{3}$$

where $\mathcal{S}$ is the set of all $L \times M$ real matrices with all columns having l_2 norm equal to $\sqrt{L}$.

3 Characterization of Sum Capacity

In this section, we shall characterize $\mathcal{C}_{\text{sum}}$. In [3] an upper bound was derived for $\mathcal{C}_{\text{sum}}$, this upper bound being the sum capacity of the "unrestricted" S-CDMA channel, i.e., the situation of no spreading when L = 1 with the appropriate power constraint on its input (sum of the power constraints of the users given by p_{tot}). The latter channel is just the M-user Gaussian multiple-access channel and its sum capacity is $C = \frac{1}{2} \log \left(1 + \frac{p_{tot}}{\sigma^2} \right)$ (see Chapter 7 in [1]) in bits/chip. When the input-power constraints are equal, it was shown in [3] that WBE signature sequences (so called because they meet the Welch-Bound-Equality (see [8])) maximize the sum capacity, and the sum capacity then equals C. However, we will show that this bound is not tight for arbitrary values of the average input-power constraints of the users and indeed there is a strict loss in sum capacity when the power constraints are "far apart"; we shall make this notion precise.

When $M \leq L$, it is easy to verify that the sum capacity is maximized when the signature sequences are chosen orthogonal to each other and we have $\mathcal{C}_{\text{sum}} = \frac{1}{2L} \sum_{i=1}^{M} \log \left(1 + \frac{L p_i}{\sigma^2} \right)$. Note that when $M = L$ and all the power constraints p_i are the same, $\mathcal{C}_{\text{sum}} = \frac{1}{2} \log \left(1 + \frac{p_{tot}}{\sigma^2} \right)$ the same as the sum capacity of the system with no spreading; this is the well known fact that for equal-power users, orthogonal multiple access incurs no loss in capacity relative to unconstrained multiple access. When $M = L$ and there is an asymmetry in the power constraints of the users, a simple argument shows that $\mathcal{C}_{\text{sum}} < C$. When

$M < L$ the claim is that $\mathcal{C}_{\text{sum}} < C$. To see this, observe that $(p_1, \ldots, p_M)$ majorizes the vector $\left(\frac{p_{tot}}{M}, \ldots, \frac{p_{tot}}{M}\right)$ (see Appendix A for the notation). It can be verified that the map $(p_1, \ldots, p_M) \mapsto \frac{1}{2L}\sum_{i=1}^{M} \log\left(1 + \frac{Lp_i}{\sigma^2}\right)$ is Schur-concave (see Appendix A for the notation and details). Hence,

$$\begin{aligned} \mathcal{C}_{\text{sum}} &= \frac{1}{2L}\sum_{i=1}^{M} \log\left(1 + \frac{Lp_i}{\sigma^2}\right) \\ &\leq \frac{M}{2L}\log\left(1 + \frac{Lp_{tot}}{M\sigma^2}\right) \\ &< \frac{1}{2}\log\left(1 + \frac{p_{tot}}{\sigma^2}\right) \end{aligned}$$

where in the last step we used the inequality $(1+x)^a < 1 + ax$ for $x > 0$ and $a \in (0,1)$.

For the case $M > L$, $\mathcal{C}_{\text{sum}}$ is characterized below. This is our main result:

Theorem 3.1 *Let $M > L$. Let (without loss of generality) $p_1 \geq \ldots \geq p_M$. Then,*

$$\mathcal{C}_{sum} = \begin{cases} \frac{1}{2}\log\left(1 + \frac{p_{tot}}{\sigma^2}\right) & \text{if } p_{tot} \geq Lp_1 \\ \frac{L-k}{2L}\log\left(1 + \frac{L\sum_{i=k+1}^{M} p_i}{(L-k)\sigma^2}\right) + \frac{1}{2L}\sum_{i=1}^{k}\log\left(1 + \frac{Lp_i}{\sigma^2}\right) & \text{if for some } 1 \leq k < L \\ \qquad \text{we have } \sum_{j=1}^{k-1} p_j + (L-k+1)\,p_k > p_{tot} \geq \sum_{j=1}^{k} p_j + (L-k)\,p_{k+1} & \end{cases} \tag{4}$$

Proof First observe that $\mathcal{C}_{\text{sum}}$ is well defined in (4). By definition from (3),

$$\begin{aligned} \mathcal{C}_{\text{sum}} &= \max_{S \in \mathcal{S}} \max_{R \in \mathcal{C}(S)} \sum_{i=1}^{M} R_i \\ &= \max_{S \in \mathcal{S}} \frac{1}{2L}\log\left[\det\left(I + \frac{1}{\sigma^2}SDS^t\right)\right] \quad \text{from (2), also see [3]} \\ &= \max_{S \in \mathcal{S}} \frac{1}{2L}\sum_{i=1}^{L}\log\left(1 + \frac{L}{\sigma^2}\lambda_i(S)\right) \end{aligned} \tag{5}$$

where $\lambda(S) = (L\lambda_1(S), \ldots, L\lambda_L(S)) \in \mathcal{R}_+^L$ denotes the vector of eigenvalues of the matrix SDS^t. Define the convex set $\mathcal{L}$ in the positive orthant of $\mathcal{R}^L$ by

$$\mathcal{L} = \left\{(\lambda_1, \ldots, \lambda_L) \in \mathcal{R}_+^L : (\lambda_1, \ldots, \lambda_L, 0, \ldots, 0) \text{ majorizes } (p_1, \ldots, p_M)\right\}$$

We first identify the region of eigenvalues of the matrix $\frac{1}{L}SDS^t$ as S varies in $\mathcal{S}$ to be exactly $\mathcal{L}$. Formally, we claim that

$$\left\{\frac{1}{L}\lambda(S) : S \in \mathcal{S}\right\} = \mathcal{L} \tag{6}$$

First consider $S \in \mathcal{S}$. Let $\tilde{\lambda}(S) \in \mathcal{R}_+^M$ be the vector of eigenvalues of the matrix $D^{\frac{1}{2}}S^tSD^{\frac{1}{2}}$ and observe that $\tilde{\lambda}(S)$ is just the vector $\lambda(S)$ with $M - L$ appended zeros. The observation that the diagonal elements of $\frac{1}{L}D^{\frac{1}{2}}S^tSD^{\frac{1}{2}}$ are $p_1, \ldots, p_M$ coupled with an appeal to Lemma A.1, allows us to conclude that $\frac{1}{L}\lambda(S) \in \mathcal{L}$. To see the other direction, consider $\lambda = (\lambda_1, \ldots, \lambda_L) \in \mathcal{L}$. Then, by definition, the vector $(\lambda_1, \ldots, \lambda_L, 0, \ldots, 0)$ majorizes the vector $(p_1, \ldots, p_M)$. Appealing to Lemma A.2, there exists a matrix H with eigenvalues $\lambda_1, \ldots, \lambda_L, 0, \ldots, 0$ and diagonal elements $p_1, \ldots, p_M$. Let $v_1, \ldots, v_L \in \mathcal{R}^M$ be the normalized eigenvectors of H corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_L$. Let $V^t = [v_1 v_2 \ldots v_L]$. If we let Λ to be the diagonal matrix with entries $\lambda_1, \ldots, \lambda_L$, then $H = V^t \Lambda V$. Now define $S = \sqrt{L}\Lambda^{\frac{1}{2}} V D^{-\frac{1}{2}}$. Then, since the square of the l_2 norms of the columns of S are the diagonal elements of $S^t S$, we verify that $S^t S = L D^{-\frac{1}{2}} H D^{-\frac{1}{2}}$ has diagonal entries equal to L concluding that $S \in \mathcal{S}$. This completes the proof of the claim in (6).

Then the sum capacity can be rewritten as, from (5),

$$\mathcal{C}_{\text{sum}} = \max_{\lambda \in \mathcal{L}} \frac{1}{2L} \sum_{i=1}^{L} \log\left(1 + \frac{L}{\sigma^2}\lambda_i\right) \tag{7}$$

The following lemma identifies a "minimal" element in $\mathcal{L}$:

Lemma 3.1 *Let* $\lambda^* = (\lambda_1^*, \ldots, \lambda_L^*) \in \mathcal{R}_+^L$ *be*

$$\lambda^* = \begin{cases} \left(\frac{p_{tot}}{L}, \ldots, \frac{p_{tot}}{L}\right) & \textit{if } p_{tot} \geq L p_1 \\ \left(p_1, \ldots, p_k, \frac{\sum_{j=k+1}^{M} p_j}{L-k}, \ldots, \frac{\sum_{j=k+1}^{M} p_j}{L-k}\right) & \textit{if for some } 1 \leq k < L \\ \textit{we have } \sum_{j=1}^{k-1} p_j + (L-k+1)\, p_k > p_{tot} \geq \sum_{j=1}^{k} p_j + (L-k)\, p_{k+1} & \end{cases} \tag{8}$$

Then

1. $\lambda^* \in \mathcal{L}$
2. *If* $\lambda \in \mathcal{L}$ *then* λ *majorizes* λ^*.

Suppose this is true. As observed earlier, the map $(\lambda_1, \ldots, \lambda_L) \mapsto \frac{1}{2L}\sum_{i=1}^{L} \log\left(1 + \frac{L\lambda_i}{\sigma^2}\right)$ is Schur-concave. Then (4) follows from (7) by an appeal to Lemma 3.1 above and the proof is complete. We only need to prove the lemma above.

Proof of Lemma 3.1:
First, observe that λ^* is well defined by the expression in (8). It is straightforward from the definition of λ^* that $\lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_L^*$ and hence that $\lambda^* \in \mathcal{L}$. Let $\lambda = (\lambda_1, \ldots, \lambda_L) \in \mathcal{L}$ and $\lambda_{[1]}, \ldots \lambda_{[L]}$ denote the order statistics of λ (see Appendix A for the notation). By the definition of λ^* in (8), it can be verified that the following relation is true among the elements of λ^*:

$$\begin{aligned} \lambda_1^* &= \max\left\{\frac{p_{tot}}{L}, p_1\right\} \\ \lambda_{k+1}^* &= \max\left\{\frac{p_{tot} - \sum_{i=1}^{k}\lambda_i^*}{L-k}, p_{k+1} + \sum_{i=1}^{k}(p_i - \lambda_i^*)\right\} \; \forall k = 1 \ldots L-1 \end{aligned} \tag{9}$$

Hence $\forall k = 1 \ldots L-1$ we can write

$$\sum_{i=1}^{k+1} \lambda_i^* = \max\left\{\sum_{i=1}^{k+1} p_i, \frac{p_{tot}}{L-k} + \frac{L-k-1}{L-k}\sum_{i=1}^{k}\lambda_i^*\right\} \tag{10}$$

Now, since $\lambda \in \mathcal{L}$ we have $\sum_{i=1}^{L} \lambda_i = p_{tot}$ and hence $\lambda_{[1]} \geq \frac{p_{tot}}{L}$. Furthermore, $\lambda_{[1]} \geq p_1$. Hence,

$$\lambda_{[1]} \geq \max\left\{\frac{p_{tot}}{L}, p_1\right\} = \lambda_1^*$$

We shall complete the proof of the claim that λ majorizes λ^* by induction. Suppose $\sum_{i=1}^{k} \lambda_{[i]} \geq \sum_{i=1}^{k} \lambda_i^*$ for some $1 \leq k < L$. Since $\sum_{i=1}^{L-k} \lambda_{[k+i]} = p_{tot} - \sum_{i=1}^{k} \lambda_{[i]}$ and $\lambda_{[k+1]} \geq \lambda_{[k+2]} \geq \ldots \geq \lambda_{[L]}$, we have $\lambda_{[k+1]} \geq \frac{p_{tot} - \sum_{i=1}^{k} \lambda_{[i]}}{L-k}$. Hence

$$\begin{aligned} \sum_{i=1}^{k+1} \lambda_{[i]} &\geq \frac{p_{tot}}{L-k} + \left(\frac{L-k-1}{L-k} \right) \sum_{i=1}^{k} \lambda_{[i]} \\ &\geq \frac{p_{tot}}{L-k} + \left(\frac{L-k-1}{L-k} \right) \sum_{i=1}^{k} \lambda_i^* \text{ by induction hypothesis} \end{aligned} \tag{11}$$

Since $\sum_{i=1}^{k+1} \left(\lambda_{[i]} - p_i \right) \geq 0$, from (11), we have

$$\begin{aligned} \sum_{i=1}^{k+1} \lambda_{[i]} &\geq \max \left\{ \sum_{i=1}^{k+1} p_i, \frac{p_{tot}}{L-k} + \frac{L-k-1}{L-k} \sum_{i=1}^{k} \lambda_i^* \right\} \\ &= \sum_{i=1}^{k+1} \lambda_i^* \quad \text{from (10)} \end{aligned}$$

This is true for all $k = 1 \ldots L-1$. Hence λ majorizes λ^* and λ^* is a Schur-minimal element of $\mathcal{L}$. This completes the proof of the lemma and hence that of the theorem. ■

As mentioned earlier, a trivial upper bound for $\mathcal{C}_{\text{sum}}$ is $C = \frac{1}{2} \log \left(1 + \frac{p_{tot}}{\sigma^2} \right)$ the sum capacity of the channel with "no spreading". We saw that when $M = L$ and all the power constraints are equal, then $\mathcal{C}_{\text{sum}}$ actually equaled C and was strictly less than C with any asymmetry in the power constraints of the users. We also saw that when $M < L$ $\mathcal{C}_{\text{sum}} < C$. For the case $M > L$, we shall now identify, as a corollary of Theorem 4.1, necessary and sufficient conditions on the input-power constraints such that $\mathcal{C}_{\text{sum}}$ will equal this upper bound. We shall identify the signature sequences for the users to achieve this sum capacity in the next section.

Corollary 3.1 *Let $M > L$. Then $\mathcal{C}_{sum} = C = \frac{1}{2} \log \left(1 + \frac{p_{tot}}{\sigma^2} \right)$ if and only if $p_{tot} \geq L p_i$ for all $i = 1 \ldots M$.*

The proof is obvious from Theorem 3.1. Theorem 3.1 also shows that when $p_{tot} < L p_i$ for some user i, then the user power constraints are not "close enough" and the resulting sum capacity of the system $\mathcal{C}_{\text{sum}}$ is strictly smaller than C. This observation follows from the fact that the map $(\lambda_1, \ldots, \lambda_L) \mapsto \frac{1}{2L} \sum_{i=1}^{L} \log \left(1 + \frac{L \lambda_i}{\sigma^2} \right)$ is strictly Schur-concave (see Appendix A for the notation and details).

4 Construction of Optimal Sequences

In this section we shall briefly comment on the construction of signature sequences that achieve sum capacity given by Theorem 3.1. We shall first focus on the case considered in Corollary 3.1. This is the situation when $\mathcal{C}_{\text{sum}}$ is equal to C, the capacity of the system with no spreading. Following the proof of Theorem 3.1, we observe that the sequences (indicated by the matrix S) for which sum capacity is achieved are precisely those with the property $SDS^t = p_{tot} I$ where, as before, D is a diagonal matrix with diagonal entries $p_1, p_2, \ldots, p_M$. The construction of such a sequence matrix $S \in \mathcal{S}$ is undertaken in [7] which gives an algorithmic procedure to construct S and illustrates the algorithm with some examples. This method is closely related to the question of constructing a symmetric matrix with a given diagonal and given eigenvalues (also see Theorem 9.B.2, [2] for the classical proof).

In the special case when all the power-constraints are equal (to say p), then $p_{tot} = Mp \geq Lp$ and hence $\mathcal{C}_{\text{sum}} = \frac{1}{2}\log\left(1+\frac{Mp}{\sigma^2}\right)$. Furthermore, the sequence matrix S constructed above from H now satisfies the relation $SS^t = MI$. This result was observed in [3] and the sequences that meet this constraint were denoted WBE sequences. Our algorithm specialized to this situation constructs WBE sequences for arbitrary $M \geq L$. A different (simpler, but still recursive) construction scheme to construct WBE sequences is in [6] where WBE sequences turn out to be optimal in a completely different context of maximizing the signal-to-interference ratio of the users in a single cell power controlled CDMA system.

Now we shall focus on the general situation when $p_{tot} \geq Lp_i$ for each i need not be satisfied. Without loss of generality, we retain the assumption that $p_1 \geq p_2 \geq \ldots \geq p_M$. Let $e_i \in \mathcal{R}^L$ be the vector $(0,0,\ldots,0,1,0,\ldots,0)$ with the entry 1 being in the ith position. Then $e_1,\ldots,e_L$ form an orthonormal basis for $\mathcal{R}^L$. Suppose the power constraints satisfy

$$\sum_{j=1}^{k-1} p_j + (L-k+1)\,p_k > p_{tot} \geq \sum_{j=1}^{k} p_j + (L-k)\,p_{k+1} \tag{12}$$

for some $k \in \{1,2,\ldots,L-1\}$. Observe that (12) is always true for some unique (depending on the power constraints) k if $p_{tot} < Lp_1$. Then, for $i=1,\ldots,k$ let $s_i = \sqrt{L}e_i$. We shall now choose the signature sequences $s_{k+1},\ldots,s_M$ for the remaining $M-k$ users from the subspace spanned by $\{e_{k+1},\ldots,e_L\}$ which has dimension $L-k$. Since $\sum_{i=k+1}^{M} p_i \geq (L-k)\,p_{k+1}$ from (12), we can appeal to the algorithm mentioned previously (see [7] and Theorem 9.B.2 in [2]) to construct sequences $s_{k+1},\ldots,s_M$ that have the property

$$\sum_{i=k+1}^{M} p_i s_i s_i^t = \left(\sum_{i=k+1}^{M} p_i\right)\begin{bmatrix} 0 & 0 \\ 0 & I_{L-k}\end{bmatrix}$$

where I_{L-k} is the identity matrix of dimension $(L-k)\times(L-k)$. It is easily verified that with this choice of signature sequences the sum capacity (given in (4))

$$\mathcal{C}_{\text{sum}} = \frac{L-k}{2L}\log\left(1+\frac{L\left(\sum_{i=k+1}^{M} p_i\right)}{(L-k)\,\sigma^2}\right) + \frac{1}{2L}\sum_{i=1}^{k}\log\left(1+\frac{Lp_i}{\sigma^2}\right)$$

is achieved.

5 Summary and Conclusion

We have completely characterized the sum capacity of a multi-user synchronous CDMA system. This characterization allowed us to derive necessary and sufficient conditions on the power constraints of the users so that for some choice of signature sequences the sum capacity of the system equals that of the system with "no spreading", namely $C = \log\left(1+\frac{p_{tot}}{\sigma^2}\right)$. We also identified the signature sequences that achieved sum capacity. A byproduct of the construction scheme is the following simple summarizing interpretation: The only assumption we retain is that the power constraints of the M users satisfy $p_1 \geq p_2 \geq \cdots \geq p_M$.

1. If $M \leq L$ the sum capacity is $\mathcal{C}_{\text{sum}} = \frac{1}{2L}\sum_{i=1}^{M}\log\left(1+\frac{Lp_i}{\sigma^2}\right)$. The choice of orthogonal signature sequences for the users is optimal and achieves sum capacity. Unless $M = L$ and $p_1 = p_2 = \cdots = p_M = p$, the sum capacity is strictly less than $C = \frac{1}{2}\log\left(1+\frac{p_{tot}}{\sigma^2}\right)$.

2. Let $M > L$ henceforth. If $p_{tot} \geq Lp_1$, then $\mathcal{C}_{\text{sum}} = C = \frac{1}{2}\log\left(1+\frac{p_{tot}}{\sigma^2}\right)$. The algorithm we derived in Section 4 can be used to construct signature sequences which achieve sum capacity.

3. Suppose $Lp_1 > p_{tot}$. Then we let user 1 have a channel of its own (we do this by letting s_1 orthogonal to all the other signature sequences) and then reduce the problem to $M-1$ users in a system with processing gain $L-1$. The resulting sum capacity is strictly smaller than C.

The allocation of signature sequences so as to achieve sum capacity emphasizes the "unfairness" of the performance criterion $\mathcal{C}_{\text{sum}}$. When $\sum_{i=1}^{L-2} p_i + 2p_{L-1} > p_{tot}$ the first $L-1$ users (these users have the largest values of p_i and hence the weakest power constraints) are allocated orthogonal signature sequences (this corresponds to allocating independent channels and are hence allowed to transmit at higher rates) while the other M-L+1 users share a common channel. In a wireless system, some users might be far away from the base station and their received powers will be correspondingly lower than users which are close to the base station. It is desirable in a practical system that resource allocation based on some choice of performance criterion of the system be "fair" to the users. Hence it is important to address fairly the situation of asymmetric power constraints on the users. One way to do this is to change the performance criterion to a weighted linear combination of the rates of the users, the weights in inverse proportion to the path gains of the users to the base station. A second way is to consider the "symmetric capacity" defined in [3]. The symmetric capacity is the sum rate of the maximum achievable equal-rate point in the union capacity region $\mathcal{C} = \bigcup_{S \in \mathcal{S}} \mathcal{C}(S)$. These questions will be answered if the characterization of the entire union capacity region $\mathcal{C}$ is done. We currently have preliminary results towards solving this open question.

A Definitions and Relevant Results from Theory of Majorization

In this appendix we collect together relevant definitions and results from the theory of majorization. All of these results can be found in [2], the comprehensive reference on majorization. Majorization makes precise the vague notion that the components of a vector x are "less spread out" or "more nearly equal" than are the components of a vector y by the statement x is majorized by y.

For any $x = (x_1, \ldots, x_n) \in \mathcal{R}^n$, let

$$x_{[1]} \geq \cdots \geq x_{[n]}$$

denote the components of x in decreasing order. For $x, y \in \mathcal{R}^n$, define

$$x \prec y \quad \text{if} \quad \begin{cases} \sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}, & k = 1 \ldots n-1 \\ \sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]} \end{cases}$$

When $x \prec y$ say that x is majorized by y (or y majorizes x). An important though trivial example of majorization is

$$\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \prec (a_1, \ldots, a_n) \tag{13}$$

for every $a \in \mathcal{R}^n$ such that $\sum_{i=1}^n a_i = 1$. An important characterization of majorization is the result that $x \prec y$ if and only if there exists a doubly stochastic matrix P such that $x = yP$.

A real valued function $\phi : \mathcal{R}^n \to \mathcal{R}$ is said to be Schur-concave if for all $x, y \in \mathcal{R}^n$ such that $x \prec y$ we have $\phi(x) \geq \phi(y)$. ϕ is said to be strictly Schur-concave if $x \prec y$ implies $\phi(x) > \phi(y)$. ϕ is said to be Schur-convex if $-\phi$ is Schur-concave. Using the observation in (13), for any Schur-concave function ϕ and for any vector $x \in \mathcal{R}^n$

$$\phi(\bar{x}) \geq \phi(x)$$

where $\bar{x} = \left(\frac{\sum_{i=1}^n x_i}{n}, \frac{\sum_{i=1}^n x_i}{n}, \ldots, \frac{\sum_{i=1}^n x_i}{n}\right)$. A well known structure of Schur-convex functions is the following result (Theorem 3.C.1 in [2]): If $g : \mathcal{R} \to \mathcal{R}$ is convex then the symmetric convex function $\phi(x) = \sum_{i=1}^n g(x_i)$ is Schur-convex. It is obvious that if $g : \mathcal{R} \to \mathcal{R}$ is concave then the symmetric concave function $\phi(x) = \sum_{i=1}^n g(x_i)$ is Schur-concave. If g is strictly concave then ϕ is strictly Schur-concave.

It is well known that the sum of diagonal elements of a matrix is equal to the sum of its eigenvalues. When the matrix is symmetric the *precise* relationship between the diagonal elements and the eigenvalues is that of majorization: (Theorem 9.B.1, [2]).

Lemma A.1 *Let H be a $n \times n$ symmetric matrix with diagonal elements $h_1, \ldots, h_n$ and eigenvalues $\lambda_1, \ldots, \lambda_n$. Then $h \prec \lambda$*

That h and λ cannot be compared by an ordering stronger than majorization is the consequence of the following converse (Theorem 9.B.2, [2]):

Lemma A.2 *If $h_1 \geq \cdots \geq h_n$ and $\lambda_1 \geq \cdots \lambda_n$ are $2n$ numbers satisfying $h \prec \lambda$ in $\mathcal{R}^n$, then there exists a real symmetric matrix H with diagonal elements $h_1, \ldots, h_n$ and eigenvalues $\lambda_1, \ldots, \lambda_n$.*

References

[1] T. M. Cover, and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

[2] A. W. Marshall, and I. Olkin, *Inequalities: Theory of Majorization and its applications*, Academic Press, 1979.

[3] M. Rupf, and J. L. Massey, "Optimum sequence multisets for Synchronous code-division multiple-access channels", *IEEE Transactions on Information Theory*, Vol 40, No. 4, pp. 1261-1266, July 1994.

[4] S. Verdú, "Capacity region of Gaussian CDMA channels: the symbol-synchronous case", in *Proc. 24th Allerton Conf.*, Oct. 1986, pp. 1025-1034.

[5] S. Verdú and S. Shamai, "Multiuser detection with random spreading and error-correction codes", *Proc. 1997 Allerton Conf. Communications, Control and Computing*, Sept-Oct 1997.

[6] P. Viswanath, V. Anantharam and D. Tse, "Optimal sequences, power control and capacity of spread-spectrum systems", College of Engg., UC Berkeley, UCB/ERL Memo No. 98/6.

[7] P. Viswanath and V. Anantharam, "Optimal sequences and sum capacity of synchronous CDMA systems", College of Engg., UC Berkeley, UCB/ERL Memo No. 98/10.

[8] L. R. Welch, "Lower bounds on the maximum cross correlation of signals", *IEEE Transactions on Information Theory*, vol.IT-20, pp.397-399, May, 1974.

Admissibility and Network Capacity of Power-Controlled CDMA Systems in Fading Channels

Junshan Zhang Edwin K.P. Chong*

School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907-1285
{junshan, echong}@ecn.purdue.edu

Abstract

Due to the fast-growing demand for network capacity in wireless networks, the characterization of network capacity has become a fundamental and pressing issue. In this paper, we study the network capacity of single-cell synchronous CDMA systems with matched filter receivers in fading channels, assuming known distributions of received powers of mobile users. We find necessary and sufficient conditions on the admissibility for both single class and multiple class systems with random signatures. For systems with deterministic signatures, we give a necessary and sufficient condition on the admissibility for single class systems, but only a sufficient condition for multiple class systems. We identify the network capacity of single class systems for both the random signature case and deterministic signature case. The network capacity can be uniquely expressed in terms of the users' QoS requirements and the distributions of the received powers.

1 Introduction

Due to the fast-growing demand for network capacity in wireless networks, it is essential to utilize efficiently the limited resources in fading channels. The characterization of network capacity is therefore one of the most fundamental and pressing issues in wireless network research. In this paper, we consider a model for the uplink of a single-cell synchronous CDMA system in fading channels.

Much of the previous research on CDMA systems has been restricted to the physical layer, focusing mainly on the performance of linear receivers in worst-case scenarios. *Near-far resistance* and *asymptotic efficiency* have been studied as two important performance measures (see [10] for a survey of the literature). At the network layer, there have been considerable efforts on power control [3, 9, 12]. The sophisticated techniques for allocating network resources at both the physical layer and network layer aim to improve network capacity. Unlike in TDMA and FDMA systems, however, there is no clean separation between resource allocation at the physical layer and network layer in CDMA systems. A natural question to address is the performance of linear receivers in power-controlled CDMA systems. In [8], Tse and Hanly characterized the network capacity for systems with several important receivers via a notion of *effective bandwidth*,

---

*Supported in part by the National Science Foundation through grant ECS-9501652 and by the U.S. Army Research Office through grant DAAH04-95-1-0246.

assuming users have random signatures. In [11], Viswanath, Anantharam, and Tse studied the joint optimization problem of designing the multiuser receiver structure, power control, and spreading sequences, and obtained simple characterizations of the resulting network capacity in various scenarios. Both [8] and [11] assumed perfect power control in characterizing network capacity. In a practical wireless communication system, however, multipath fading is ubiquitous, making perfect power control impossible. Little work has been done on network capacity of CDMA systems with multiuser receivers in fading channels. In this paper, we study the admissibility and network capacity in power-controlled CDMA systems in fading channels. All our results assume a matched filter receiver, which is the designated receiver for CDMA systems in IS-95.

The organization of the rest of this paper is as follows. In Section 2 we give some background for our model of power-controlled CDMA systems. In Section 3, we study the admissibility and network capacity for single class systems, and define the concept of *effective target SIR.* In Section 4, we extend the study to multiple class systems, and define *effective bandwidth* for power-controlled CDMA systems in fading channels. Finally, we draw some conclusions in Section 5.

2 Preliminaries

In a synchronous CDMA system, each user, say user k, is assigned a binary (± 1 valued) signature of length L, $S_k = \frac{1}{\sqrt{L}}(W_{k1}, \ldots, W_{kL})^T$, where W_{ki} is either 1 or -1 for $i = 1, \ldots, L$. The user's information symbols are spread onto its signature. We will study CDMA systems for the following two cases. The first is the deterministic signature case, by which we mean that the users' signatures are jointly designed and deterministic when they are admitted into the system (cf., [6, 11]); the second is the random signature case, by which we mean that users choose their signatures randomly and independently when they are admitted into the system, which models the situation where the information symbols are spread onto pseudo-random spreading sequences (cf., [8]).

Suppose users transmit data over a fading channel. Typically, fading channel gains are assumed to be stationary and ergodic, and all the users in one class are assumed to have independently and identically distributed channel gains (see [5]). As pointed out in [2] and [8], it is ideal to have the received powers for all the users in one class to be the same all the time. However, it is very difficult to implement perfect power control and the received powers fluctuate around the desired levels in a practical system. Assuming that the same kind of power control algorithm is applied to all users in one class, by symmetry, we model the received powers for all users in one class to be independently and identically distributed in a large network. For technical reasons, we also assume that power control is good enough to ensure that the fluctuation around the expected received power is bounded (say by d) with probability one.

We assume that the distributions of the received powers are known. This assumption is reasonable (cf., [4, Chapter 6]). From a practical viewpoint, the distributions of the received powers can usually be obtained through measurements (see [7]). We aim to study the admissibility and network capacity in the limiting regime, that is, when both M and L go to infinity. As in [8], $\beta \equiv \frac{M}{L}$ is taken to be fixed as L goes to infinity.

3 Single Class Systems

Based on [8, 9], we have the following discrete-time model for a synchronous CDMA system. Suppose there are M users in the system when the processing gain is L. For any symbol interval, the baseband received signal at the front of the receiver is

$$Y = \sum_{k=1}^{M} \sqrt{P_k^{(L)}} b_k S_k + V, \tag{1}$$

where the b_k's are the transmitted information symbols and the $P_k^{(L)}$'s are received powers for $k = 1, \ldots, M$, and V is $\mathcal{N}(0, \eta I)$ background white noise that comes from the sampling of the ambient white Gaussian noise with power spectral density η. The received powers are assumed to be i.i.d. with distribution F_L, which has a density function and mean $\mu^{(L)}$. When L goes to infinity, we assume that the sequence of functions $\{F_L\}$ converges weakly to F, which denotes the distribution of the received powers in the limiting regime. Let P_0 denote a random variable that has distribution F and mean μ. For convenience, we further assume the modulation is antipodal, that is, $b_k \in \{-1, 1\}$.

Assume the signature S_k of user k is known to the receiver for it. Then the received SIR (signal-to-interference ratio) for user k after passing the matched filter is

$$SIR_k^{(L)} = \frac{P_k^{(L)}}{\eta + \sum_{\substack{i=1 \\ i \neq k}}^{M} P_i^{(L)} (S_k^T S_i)^2} \qquad k = 1, \ldots, M. \tag{2}$$

Since the received SIR's are random variables under our assumption, we adopt a probabilistic model for the users' QoS requirements as follows:

$$P\{SIR_k^{(L)} \geq \gamma\} \geq a \qquad \forall\, k = 1, \ldots, M.$$

3.1 The Deterministic Signature Case: Admissibility and Network Capacity

In this case, the users' signatures are deterministic, and the signatures of all the M users form a signature set when the processing gain is L. When L goes to infinity, we have a sequence of signature sets. Let $\mathcal{F}$ denote the collection of sequences of signature sets satisfying the following condition:

$$\sup_{\substack{i,j \leq M \\ i \neq j}} (S_i^T S_j)^2 \log L \to 0 \qquad \text{as } L \to \infty, \tag{3}$$

where S_i and S_j are the signatures for user i and j respectively when the processing gain is L, $i, j \in \{1, \ldots, M\}$. In the proceeding parts where sequences of deterministic signature sets are chosen, we assume that all sequences are in $\mathcal{F}$.

First we define *admissibility* for a class of users as follows: when the processing gain is L, β users per degree of freedom of the processing gain are admissible in the system if and only if there exists a signature set $\{S_1, \ldots, S_M\}$ for the M users, where M equals βL, and the users' QoS requirements are satisfied when $\{S_1, \ldots, S_M\}$ are allotted to the users, that is,

$$P\{SIR_k^{(L)} \geq \gamma\} \geq a, \quad \forall k = 1, \ldots, M,$$

which is equivalent to

$$\max_{\{S_1,\ldots,S_M\}} \min_k P\{SIR_k^{(L)} \geq \gamma\} \geq a. \tag{4}$$

Asymptotic admissibility is defined as follows: we say β users per degree of freedom of the processing gain are asymptotically admissible when the following condition holds:

$$\lim_{L\to\infty} \max_{\{S_1,\ldots,S_M\}} \min_k P\{SIR_k^{(L)} \geq \gamma\} \geq a.$$

As pointed out in [11], when $M \leq L$, we can always choose an orthogonal signature set such that there is no interference between users. Then each single user transmits data as if it were in a single-user channel. This case is not of real interest since we want to find the maximum number of users supportable by the system. For the case $M > L$, our results make use of WBE (Welch Bound Equality) signature sets. (See [6] for the definition of WBE sequence sets. Note that we use the terminology "signature set" instead of "sequence set" to avoid confusion since we will discuss sequences of such sets.)

Assuming $\lim_{L\to\infty} \max_k \sum_{j=1}^{M} (S_k^T S_j)^2 < \infty$, we have the following proposition on the asymptotic admissibility for the deterministic signature case.

Proposition 3.1 *Suppose all sequences of deterministic signature sets are chosen from $\mathcal{F}$. Then β users per degree of freedom of the processing gain are asymptotically admissible if and only if $\beta \leq \frac{F^{-1}(1-a)}{\gamma\mu} + 1 - \frac{\eta}{\mu}$.*

Proposition 3.1 illustrates the fact that the asymptotically admissible region is $\beta \leq \frac{F^{-1}(1-a)}{\gamma\mu} + 1 - \frac{\eta}{\mu}$. Following [8], we can call $\frac{F^{-1}(1-a)}{\gamma\mu} + 1 - \frac{\eta}{\mu}$ the *network capacity in the limiting regime* (to be distinguishable from the concept "asymptotic network capacity" which is introduced below) for the given distribution of the received powers.

In practical CDMA systems, the processing gain is finite and large. Therefore, one basic question to ask is, "How many users are admissible for some large (but finite) processing gain while the users' QoS requirements are fixed?" Basing only on Proposition 3.1, however, we cannot answer this interesting question. (The subtlety will be more clear in the proceeding discussions.) This motivates us to develop another approach that studies a sequence of network capacities indexed by the processing gain L. We proceed to elaborate on this idea in detail.

When the processing gain is L, we let C_L denote the set of all the admissible points, that is,

$$C_L = \left\{ \beta \,\middle|\, \max_{\{S_1,\ldots,S_M\}} \min_k P\{SIR_k^{(L)} \geq \gamma\} \geq a \right\}.$$

Of particular interest is the maximum number of users admissible by the system, which is defined to be the *network capacity* in terms of the number of users per degree of freedom of the processing gain. That is, the network capacity is the maximum number of users per degree of freedom of the processing gain that are admissible in the system, and depends on a. Therefore, when the processing gain is L, we can define *network capacity* as follows:

$$\beta_L(a) = \sup C_L = \sup \left\{ \beta \,\middle|\, \max_{\{S_1,\ldots,S_M\}} \min_k P\{SIR_k^{(L)} \geq \gamma\} \geq a \right\}. \tag{5}$$

The *asymptotic network capacity* $\beta_\infty(a)$ is defined as the limit of $\beta_L(a)$ when $L \to \infty$, that is, $\beta_\infty(a) = \lim_{L\to\infty} \beta_L(a)$. Note that the network capacity in the limiting regime and the asymptotic network capacity are different in principle (compare the two definitions). A priori, we do not know if they lead to the same expression.

We aim to identify the asymptotic network capacity. For technical reasons, we assume that the received powers in the limiting regime have a connected support, which implies that the distribution function F is strictly increasing over its support. We have the following proposition on the asymptotic network capacity for the deterministic signature case.

Proposition 3.2 *Suppose all sequences of deterministic signature sets are chosen from $\mathcal{F}$. Then the asymptotic network capacity is $\beta_\infty(a) = \frac{F^{-1}(1-a)}{\gamma\mu} + 1 - \frac{\eta}{\mu}$, which can be achieved through a sequence of WBE signature sets in $\mathcal{F}$.*

It turns out that for the deterministic signature case, the *network capacity in the limiting regime* and the *asymptotic network capacity*, which are different in principle, lead to the same result under the auxiliary assumption made on the distribution function F. In general, the network capacity in the limiting regime is an upper bound of the asymptotic network capacity. The asymptotic network capacity, however, is a more appropriate approximation of the network capacity of a large network. In particular, by exploiting Proposition 3.2, we can conclude that for a given $\epsilon > 0$, $\beta_\infty(a) - \epsilon$ users per degree of freedom of the processing gain are admissible in the system for very large L ($L \geq L_0(\epsilon)$ for some $L_0(\epsilon)$).

Worth noting is that sequences of WBE signature sets in $\mathcal{F}$ play a very important role in both the asymptotic admissibility and asymptotic network capacity. A heuristic interpretation is that the choice of WBE signature sets is asymptotically optimal in the sense of suppressing the interference of all the users simultaneously.

3.2 The Random Signature Case: Admissibility and Network Capacity

In this case, users choose signatures randomly and independently, which models the situation where the information symbols are spread onto pseudo-random spreading sequences. When the processing gain is L, the model for random signatures is as follows: $S_k = \frac{1}{\sqrt{L}}(W_{k1}, \ldots, W_{kL})^T$, where the W_{ki}'s are i.i.d. with $P\{W_{ki} = 1\} = P\{W_{ki} = -1\} = \frac{1}{2}$ for $i = 1, \ldots, L$ and $k = 1, \ldots, M$.

Since the $P_k^{(L)}$'s are i.i.d., and the W_{ki}'s are i.i.d., the received SIR's for users are also identically distributed. Hence, $P\{SIR_k^{(L)} \geq \gamma\}$ does not depend on k, and we can study user 1 without loss of generality.

Assume that the signature for a user is known to the receiver for it. Given the distribution of the received powers, we define admissibility for a class of users with random signatures as follows: when the processing gain is L, β users per degree of freedom of the processing gain are *admissible* in the system if

$$P\{SIR_1^{(L)} \geq \gamma\} \geq a.$$

We say β users per degree of freedom of the processing gain are *asymptotically admissible* if the following condition holds:

$$\lim_{L\to\infty} P\{SIR_1^{(L)} \geq \gamma\} \geq a.$$

We have the following proposition on the asymptotic admissibility for the random signature case.

Proposition 3.3 *Suppose users choose signatures randomly and independently. Then* β *users per degree of freedom of the processing gain are asymptotically admissible if and only if* $\beta \leq \frac{F^{-1}(1-a)}{\gamma\mu} - \frac{\eta}{\mu}$.

We call $\frac{F^{-1}(1-a)}{\gamma\mu} - \frac{\eta}{\mu}$ the *network capacity in the limiting regime* for the given distribution of the received powers.

Similar to the deterministic signature case, when the processing gain is L, the *network capacity* is defined as the maximum number of users per degree of freedom of the processing gain that are admissible in the system, that is,

$$\beta_L(a) = \sup\left\{\beta \,\middle|\, P\{SIR_1^{(L)} \geq \gamma\} \geq a\right\}.$$

We define *asymptotic network capacity* $\beta_\infty(a)$ as the limit of $\beta_L(a)$ when $L \to \infty$, that is, $\beta_\infty(a) = \lim_{L\to\infty} \beta_L(a)$.

Again, for technical reasons, we assume the received powers in the limiting regime have a connected support and hence F is strictly increasing over its support. Then we have the following proposition on the asymptotic network capacity for the random signature case.

Proposition 3.4 *Suppose users choose their signatures randomly and independently. Then the asymptotic network capacity is* $\beta_\infty(a) = \frac{F^{-1}(1-a)}{\gamma\mu} - \frac{\eta}{\mu}$.

It turns out that for the random signature case, the *network capacity in the limiting regime* and the *asymptotic network capacity*, which are different in principle, also lead to the same expression under the auxiliary assumption made on the distribution function F. It can easily be observed that a system using optimal deterministic signatures yields precisely one more user per degree of freedom of the processing gain than a system with signatures randomly and independently chosen, which is a generalization of the result given in [11] under perfect power control.

3.3 Effective Target SIR

We proceed to look for an abstraction that can combine the information of both QoS requirements and the distributions of received powers for determining the asymptotic admissibility. In a power-controlled system, the actual received powers depend on the fading environments and the performance of the power control algorithms. Note that if β users are not asymptotically admissible for a given received power distribution, it may still be possible to make β asymptotically admissible by scaling up the received power. Referring to the definition of *scale family* in [1, p. 118], we consider the scale family $\mathcal{G} = \{F(\frac{P_0}{c}), c > 0\}$, where F is the distribution function of P_0. Define $\gamma' = \gamma c_e$, where $c_e = \frac{\mu}{F^{-1}(1-a)}$. Clearly, c_e is fixed for the scale family $\mathcal{G}$. We call γ' the *effective target SIR* for $\mathcal{G}$.

Note that in a given scale family of distributions, there exists a one-to-one correspondence between distribution and mean. Given a scale family $\mathcal{G}$ of received power distributions, a basic question to ask is if there exists a finite positive value that can be designated as μ such that β users per degree of freedom of processing gain are asymptotically admissible for the corresponding received power distribution in $\mathcal{G}$. For the deterministic signature case, it can be shown that there exists a finite positive solution for μ if and only if $\beta\frac{\gamma'}{1+\gamma'} < 1$; moreover, the minimum value for μ is $(\eta\frac{\gamma'}{1+\gamma'}) \,/\, (1 - \beta\frac{\gamma'}{1+\gamma'})$. For the random signature case, it is easy to obtain that there exists a finite positive solution for μ if and only if $\beta\gamma' < 1$; the minimum value for μ is $\frac{\eta\gamma'}{1-\beta\gamma'}$.

4 Multiple Class Systems

We have studied the asymptotic admissibility and identified the asymptotic network capacity for single class systems in the above section. However, future wireless systems will have to support multimedia services such as voice, data, video, and fax. Therefore, it is essential to have a level of generalization dealing with users having different QoS requirements.

Suppose there are a fixed number, say N, of classes. For $n = 1, \ldots, N$, let A_n denote the set of users in class n and M_n the cardinality of A_n. Let S_{in} denote the signature for the ith user in class n. Define $\beta_n = \frac{M_n}{L}$ and let β_n be fixed when $L \to \infty$ as in the single class case.

The received SIR for the ith user in class n is

$$SIR_{in}^{(L)} = \frac{P_{in}^{(L)}}{\eta + \sum_{\substack{j \neq n \\ l \in A_j}} P_{lj}^{(L)} (S_{in}^T S_{lj})^2 + \sum_{\substack{l \in A_n \\ l \neq i}} P_{ln}^{(L)} (S_{in}^T S_{ln})^2}, \tag{6}$$

where $P_{in}^{(L)}$ denotes the received powers for $i \in A_n$ and $n = 1, \ldots, N$ when the processing gain is L.

Again our results are asymptotic in nature. For the sake of simplification, as in the single class case, we let P_n denote a random variable that has the same distribution F_n as the received powers of users in class n in the limiting regime, and assume that F_n has a density function and mean μ_n. For convenience, we call the collection $\{F_1, \ldots, F_N\}$ a *group of received power distributions*, and $\vec{P} = (\mu_1, \ldots, \mu_N)^T$ the *mean power vector.* Suppose it is required that the users in class n must have received SIR greater than or equal to γ_n with probability no less than $a_n \in [0, 1]$.

4.1 The Deterministic Signature Case: Admissibility, Effective Target SIR, and Effective Bandwidth

We assume that every chosen sequence of signature sets satisfies (3) and $\lim_{L \to \infty} \max_{i \in A_n} \sum_{l \in A_j} (S_{in}^T S_{lj})^2 < \infty$ for all $n,\ j \in \{1, \ldots, N\}$. Let $I_{in} = S_{in}^T D S_{in}$, where $D = \sum_{j=1}^{N} \sum_{l \in A_j} S_{lj} S_{lj}^T \mu_j$. Then we have

$$SIR_{in}^{(L)} \xRightarrow{\mathcal{D}} \frac{P_n}{\eta + I_{in} - \mu_n} \qquad \text{as } L \to \infty.$$

Similar to the argument in the single class case, it is desirable to design the signature sets such that all users' QoS requirements are satisfied. When $\sum_{n=1}^{N} \beta_n \leq 1$, orthogonal signature sets will work and each single user transmits data as if it were in a single-user channel. When $\sum_{n=1}^{N} \beta_n > 1$, however, a simple closed-form solution to the global optimization of the signature sets for this case seems unattainable. Therefore, we study a sufficient condition for the asymptotic admissibility of multiple class systems for this case.

For the sake of simplification, first we let $F_{M,n}$ denote the signature set of class n when the processing gain is L. Given the distributions of the received powers, we say a tuple $(\beta_1, \ldots, \beta_N)$ is asymptotically admissible if the following condition holds:

$$\lim_{L \to \infty} \max_{\cup_n F_{M,n}} \min_{i \in A_n} P\{SIR_{in}^{(L)} \geq \gamma\} \geq a_n \qquad \forall\, n = 1, \ldots, N.$$

We have the following proposition on the asymptotic admissibility of users of multiple classes for the deterministic signature case.

Proposition 4.1 *A tuple $(\beta_1, \ldots, \beta_N)$ is asymptotically admissible if it satisfies the following two conditions:*

$$\max_n \mu_n \;\leq\; \sum_{j=1}^{N} \beta_j \mu_j, \tag{7a}$$

$$\sum_{j=1}^{N} \beta_j \mu_j \;\leq\; \min_n \left(\frac{F_n^{-1}(1-a_n)}{\gamma_n} + \mu_n \right) - \eta. \tag{7b}$$

As in the single class case, we introduce the scale family $\mathcal{G}_n = \{F_n(\frac{p_n}{c}), c > 0\}$ for P_n, where F_n is the distribution function of P_n, $n = 1, \ldots, N$. Define $\gamma_n' = \gamma_n c_{en}$, where $c_{en} = \frac{\mu_n}{F_n^{-1}(1-a_n)}$. As before, c_{en} is fixed for the scale family $\mathcal{G}_n$. We call γ_n' the *effective target SIR* for $\mathcal{G}_n$. For ease of reference, we call the collection $\{\mathcal{G}_1, \ldots, \mathcal{G}_N\}$ a *group of scale families.*

Observe that in a given group of scale families, there exists a one-to-one correspondence between the mean power vector and the group of received power distributions. We have the following result for a given group of scale families $\{\mathcal{G}_1, \ldots, \mathcal{G}_N\}$.

Proposition 4.2 *There exists a finite positive vector that can be assigned as the mean power vector for a tuple $(\beta_1, \ldots, \beta_N)$ to be asymptotically admissible for the corresponding group of received power distributions in $\{\mathcal{G}_1, \ldots, \mathcal{G}_N\}$ if*

$$\frac{\gamma_n'}{1+\gamma_n'} \leq \sum_{n=1}^{N} \beta_n \frac{\gamma_n'}{1+\gamma_n'},$$

and

$$\sum_{n=1}^{N} \beta_n \frac{\gamma_n'}{1+\gamma_n'} < 1.$$

Observing the conditions given in Proposition 4.2, we are motivated to extend the insightful idea in [8] and define the *effective bandwidth* of class n for the deterministic signature case as $e(\gamma_n) = \frac{\gamma_n'}{1+\gamma_n'}$ degrees of freedom per user. However, we are able to give only a sufficient condition for asymptotical admissibility of a tuple $(\beta_1, \ldots, \beta_N)$ in terms of effective bandwidth. Further work is needed for this case.

4.2 The Random Signature Case: Admissibility, Effective Target SIR, and Effective Bandwidth

Similar to the approach in the single class systems with random signatures, we get

$$SIR_{in}^{(L)} \overset{\mathcal{D}}{\Longrightarrow} \frac{P_n}{\eta + \sum_{n=1}^{N} \beta_n \mu_n} \qquad \text{as } L \to \infty. \tag{8}$$

We can study the first user in class n without loss of generality.

Given the distributions of the received powers, we say a tuple $(\beta_1, \ldots, \beta_N)$ is asymptotically admissible if the following condition holds:

$$\lim_{L\to\infty} P\{SIR_{1n}^{(L)} \geq \gamma\} \geq a_n \qquad \forall\, n = 1, \ldots, N.$$

We have the following proposition on the asymptotic admissibility of users of multiple classes for the random signature case.

Proposition 4.3 *Suppose users choose their signatures randomly and independently. A tuple* $(\beta_1, \ldots, \beta_N)$ *is asymptotically admissible if and only if*

$$\sum_{n=1}^{N} \beta_n \mu_n \leq \min_n \left(\frac{F_n^{-1}(1-a_n)}{\gamma_n} \right) - \eta. \tag{9}$$

Again we define $\gamma_n' = \gamma_n c_{en}$, where $c_{en} = \frac{\mu_n}{F_n^{-1}(1-a_n)}$ is fixed for the scale family $\mathcal{G}_n$, $n = 1, \ldots, N$. We can easily obtain the following result for a given group of scale families $\{\mathcal{G}_1, \ldots, \mathcal{G}_N\}$.

Proposition 4.4 *a). There exists a finite positive vector that can be assigned as the mean power vector for a tuple* $(\beta_1, \ldots, \beta_N)$ *to be asymptotically admissible for the corresponding group of received power distributions in* $\{\mathcal{G}_1, \ldots, \mathcal{G}_N\}$ *if and only if*

$$\sum_{n=1}^{N} \beta_n \gamma_n' < 1.$$

Similar to the deterministic signature case, we define the *effective bandwidth* of class n for the random signature case as $e(\gamma_n) = \gamma_n'$ degrees of freedom per user. Proposition 4.4 states that for a given group of scale families, a tuple $(\beta_1, \ldots, \beta_N)$ can be made asymptotically admissible through power control if and only if the sum of effective bandwidth of all classes is less than one.

5 Conclusions

We have studied single-cell synchronous CDMA systems with matched filter receivers assuming known distributions of received powers of mobile users. We studied the asymptotic admissibility for both single class systems and multiple class systems. For systems with random signatures, we gave necessary and sufficient conditions on the asymptotic admissibility in terms of effective bandwidth. For users with deterministic signatures, we gave a necessary and sufficient condition for single class systems, but only a sufficient condition for multiple class systems. Further work is needed for this case.

We have also identified the asymptotic network capacity of single class systems for both the deterministic signature case and the random signature case. The asymptotic network capacity for single class systems can be uniquely expressed in terms of the users' QoS requirements and the distributions of the received powers. We also identified the optimal sequences of signature sets for the deterministic signature case in an asymptotic setting. Furthermore, we explored the concepts of effective target SIR and effective bandwidth, which play an important role in determining the asymptotic admissibility and hence the asymptotic network capacity. Our results are useful for network-level resource allocation problems such as admission control and power control in a large network.

In this work we have focused on finding the network capacity so that we can determine how many users can be accommodated without sacrificing their QoS requirements. On the other hand, another fundamental issue is the *channel capacity.* Currently we are looking into this problem.

References

[1] G. Casella and R. L. Berger, *Statistical Inference.* Duxbury Press, 1990.

[2] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver Jr., and C. E. Wheatley III, "On the capacity of a cellular CDMA system," *IEEE Transactions on Vehicular Technology*, vol. 40, pp. 303–312, May 1991.

[3] S. V. Hanly, "Capacity and power control in spread spectrum macro-diversity radio networks," *IEEE Transactions on Communication*, vol. 4, no. 2, pp. 247–256, 1996.

[4] S. V. Hanly, *Information Capacity of Radio Networks.* PhD thesis, Cambridge University, Aug. 1993.

[5] S. V. Hanly and D. Tse, "Multi-access fading channels: Shannon and delay-limited capacities," in *Proc. 33rd Allerton conf.*, pp. 786–795, Oct. 1995.

[6] J. L. Massey and T. Mittelholzer, "Welch's bound and sequence sets for code-division multiple-access systems," in *Sequences II, Methods in Communications, Security and Computer science* (R. Capocelli, A. D. Santis, and U. Vaccaroo, eds.), Springer Verlag, 1993.

[7] Z. Rosberg and J. Zander, "Toward a framework for power control in cellular systems," *Wireless networks*, vol. 4, no. 3, pp. 215–222, Apr. 1998.

[8] D. Tse and S. V. Hanly, "Multiuser demodulation: Effective interference, effective bandwidth and capacity," preprint, Dec. 1997.

[9] S. Ulukus and R. D. Yates, "Stochastic power control for cellular radio systems," *IEEE Transactions on Communication*, vol. 46, no. 6, pp. 784–798, 1998.

[10] S. Verdú, "Multiuser detection," in *Advances in Statistical Signal Processing*, vol. 2, pp. 369–409, JAI Press Inc., 1993.

[11] P. Viswanath, V. Anantharam, and D. Tse, "Optimal sequences, power control and capacity of spread-spectrum systems with multiuser receivers," preprint, Jan. 1998.

[12] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341– 1348, Sep. 1995.

2D-RAKE combiner for DS-CDMA systems in the presence of pilot transmissions

Jinho Choi and Gi Hun Lee
Advanced Wireless Tech. Lab.
LG Information and Communications
{jin, ghl}@lgic.co.kr

Abstract

In the presence of pilot transmissions in the reverse-link, two approaches to compute the weight vectors for the 2D-RAKE combiner of base-station receivers are proposed. In the approaches, since the weight vectors are computed from the despread signal, the computational complexity is relatively lower than the approaches which use the received signal before despreading.

1 Introduction

Recently, third generation wireless communication systems are proposed for higher bit rate services [1] with code division multiple access (CDMA). A feature of those systems is the coherent reverse-link. It is claimed that the coherent reverse-link approach provides better performance than the noncoherent reverse-link approach. An example of the mobile station transmitters for third generation systems is shown in Figure 1. For the coherent reverse-link, there is a pilot transmission associated with the traffic transmission. Note that the power of the pilot transmission would be lower than that of the traffic transmission to reduce the effect of the interference due to pilot transmissions from other users.

In order to provide better link quality and/or higher capacity, recently multiple antennas are considered in base-station receivers [2]. For third generation CDMA systems with antenna arrays, it would be required to develop algorithms for the space-time modems in the presence of pilot transmissions.

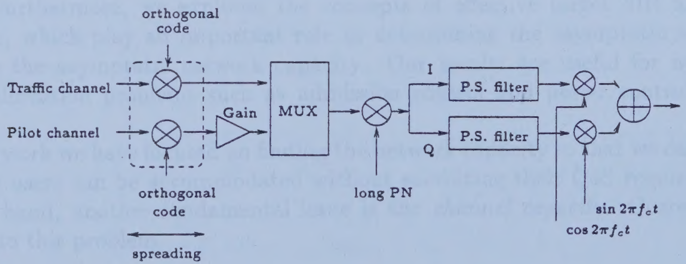


Figure 1: An example of transmitter of a mobile station for coherent reverse-link.

2 Approaches and discussion

We propose two approaches to compute the weight vectors (beamforming vectors) for the 2D-RAKE combiner in the presence of pilot transmissions. The approach in [2] is modified to utilize pilot transmissions. In the approach in [2], the covariance matrices of the received signal vectors before despreading are required to compute the weight vectors. In order to compute these covariance matrices, a number of multiplications are required at the chip rate. For a higher chip rate, those approaches obviously would not be practical because of high computational complexity. To overcome the above difficulty, the proposed approaches in this paper provide the weight vectors from the despread signal vectors. Thus, all operations for computing the weight vectors are processed at the symbol rate that is generally much lower than the chip rate. Thus, the proposed approaches can provide the weight vectors with relatively lower computational complexity.

The channel vectors can be directly estimated from the pilot transmission. The approach which estimates the channel vector from the pilot transmission is referred to as the direct approach. By using both traffic and pilot transmissions, the channel vectors can be obtained. In this case, the eigendecomposition of the covariance matrices is required. This approach is referred to as the eigendecomposition approach. In Figure 2, the SINR curves are shown for various values of the pilot to traffic power ratio (PTR). For a lower PTR, the eigendecomposition approach is superior to the direct approach. For a higher PTR, the direct approach is better. Note that as the PTR gets higher, the theoretical (attainable) SINR gets lower. Thus, for better link quality or capacity, the PTR shall be low and the eigendecomposition approach would be preferable.

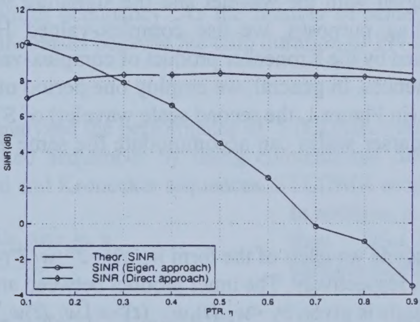


Figure 2: SINR versus PTR.

References

[1] TR45.5.4 RTT, "Working draft for ITU-R RTT candidate submission (V0.11)," February 1998.

[2] A.F. Naguib, *Adaptive antennas for CDMA wireless networks*, PhD thesis, Stanford University, 1996.

PERFORMANCE OF SCALE-TIME-CODE DIVISION MULTIPLE ACCESS FOR COMPLEX-VALUED PN SEQUENCES

Oguz Kucur and Guillermo E. Atkin
Illinois Institute of Technology, ECE Dept., 3301 S. Dearborn St., Chicago IL 60616
E-mail: atkin@charlie.iit.edu, Tel: (312) 567 6810

I. Introduction

We have recently proposed a new multiple access system called ***Scale-Time-Code-Division Multiple Access*** (**STCDMA**) which depends on the time, code, and scale orthogonality introduced by wavelets [1], [2]. We use wavelets as an orthonormal set of symbols for signaling, and exploit their orthonormality over scale and time. In a STCDMA system, the channel is partitioned into different scales, and each scale into different time slots. For example, in the dyadic case (scale parameter is a power of *2*) the first scale (the coarsest one) uses only one time slot *[0, T]*, using the basic wavelet. The following finer scale uses two time slots *[0, T/2]*, *[T/2, T]*; the third one four time slots and so on. In addition, pseudo-noise (PN) sequences are used in each time slot to identify multiple users. Each transmitter uses a specific scale and time slot and is assigned a distinct PN sequence of different length that fits its corresponding time slot. So each user encodes its information symbols with a time-shifted and scaled replica (appropriate for its assigned scale and time slot) of the same basic wavelet, and then spreads its translated and scaled wavelet by its PN sequence. At the receiver, the information streams for each user are decoded by using the orthogonality of the wavelets and of the PN sequences.

Figure 1 shows *2*-scale STCDMA format using dyadically scaled Haar wavelet. We will use antipodal signaling at the transmitter. At the receiver, we use a conventional detector for each user due to the orthogonality of wavelets and of the PN sequences. The conventional detector correlates the received signal with the wavelet and the signature waveform assigned by the desired user. For spreading purposes, we use complex-valued Hadamard sequences and Kronecker sequences formed by the Kronecker product of complex-valued Hadamard sequences and real-valued Gold sequences. In general, we employ one period of the signature waveforms in the finest scale wavelet (in Figure 1, the second scale wavelet) of STCDMA so that the users in the time slots of the coarser scales can accommodate the same signature waveforms with multiple periods.

II. Wavelets

In our work, we use discrete wavelets of the form $w_{m,n}(t)=2^{m/2}w(2^m t-nT)$, where m, n are scale and translation parameters, respectively. The inner product between any two wavelets $w_{k,l}(t)$ and $w_{m,n}(t)$ in an orthonormal basis is given by $<w_{k,l}(t),w_{m,n}(t)>=\int w_{k,l}(t)w^{*}_{m,n}(t)dt=\delta_{k,m}\delta_{l,n}$, where $\delta_{i,j}$ is the Kronecker delta function.

III. Complex-valued PN sequences

In this work, we will use complex-valued Hadamard sequences and Kronecker sequences formed by the Kronecker product of complex-valued Hadamard sequences and real-valued Gold sequences. Hadamard codewords are the rows of Hadamard matrix. We define a complex-valued Hadamard matrix $\hat{H}_N=[\hat{h}_i(k)]$ of order N (N is a power of 2) as follows

$$\hat{H}_N=\begin{bmatrix} H_{N/2} & H_{N/2} \\ jH_{N/2} & -jH_{N/2} \end{bmatrix} \qquad (1)$$

where $j=\sqrt{-1}$ and $-H_{N/2}$ is the complement of the matrix $H_{N/2}$. In Equation (1), $H_{N/2}$ is a real-valued Hadamard matrix of order $N/2$. In general, a real-valued Hadamard matrix H_N of order

N is an NxN matrix with entries ∓ 1 such that the rows are orthogonal. It can be shown that the orthogonality property is also preserved for the complex-valued Hadamard matrices.

IV. Performance analysis

The total STCDMA signal of 2-scale format for synchronous transmission is described as

$$s(t)=\sum_{m=0}^{1}\sum_{n=0}^{2^m-1}\sum_{i=1}^{L}\sqrt{2\epsilon}\; b_{m,n,i}\; w_{m,n}(t)\;\cos\{\omega_c(t-nT/2^m)+r_{il}(t-nT/2^m)\} \qquad 0\le t\le T \qquad (2)$$

where $b_{m,n,i}$ is the information bit for the user transmitting at the nth time slot of the mth scale (there are 2^m time slots in each scale) using the ith quadriphase signature waveform of l periods, $r_{il}(t)$, and ω_c, L, T are the carrier frequency, the number of PN sequences and the symbol interval in the first scale, respectively. ϵ is the signal power for each user (power control is exercised). The transmitted signal is assumed to be corrupted by additive white Gaussian noise (AWGN) with power spectral density $N_0/2$. In addition, we assume that time and phase synchronism are maintained. At the receiver site, each user employs conventional detector. Each user's matching waveform is $w_{m,n}(t)cos\{\omega_c(t-nT/2^m)+r_{il}(t-nT/2^m)\}$.

In the analysis, for complex-valued Hadamard and Kronecker sequences the multiple access interference from the desired user's slot and from other scales vanish and the probability of bit error is obtained as $P_e=0.5erfc\{\sqrt{\gamma_b}\}$, i.e., the performance of BPSK. Therefore, for these sequences, STCDMA supports more users than DS-CDMA for the same bandwidth occupancy and the same BPSK performance if the bottom scale is thought to be DS-CDMA with time-division multiplexing. For example, 3-scale STCDMA supports $7N$ users as compared to $4N$ users of DS-CDMA, which is a very important capacity improvement if the third scale of 3-scale STCDMA format is thought to be DS-CDMA with time division multiplexing. Although STCDMA seems to support as many users as desired, the number of users supported is constrained by the bandwidth occupancy. As the number of scale increases, the bandwidth occupancy of the system increases because the chip duration of PN sequences decreases.

V. Conclusions

In this work, we have analyzed the performance of STCDMA over the synchronous AWGN channel for complex-valued sequences by using conventional detector for each user. For complex-valued Hadamard and Kronecker sequences, STCDMA achieves BPSK performance. In addition, it supports much more users than DS-CDMA for the same performance and bandwidth occupancy.

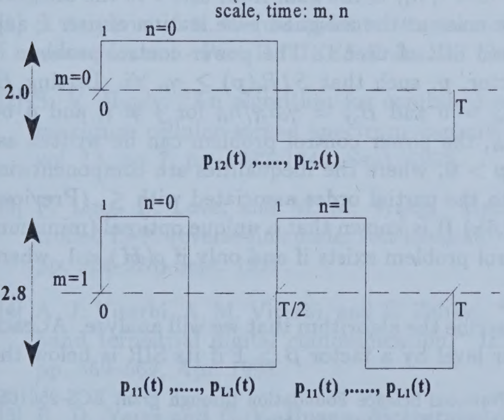


Figure 1. 2-scale STCDMA format

REFERENCES

[1] O. Kucur and G. E. Atkin, "Scale-Time-Code Division Multiple Access (STCDMA)," *Proc. of the 1996 Tactical Comm. Conf.*, Fort Wayne, pp. 449-457, April 1996.

[2] O. Kucur and G. E. Atkin, "A Wavelet-Based Code Division Multiple Access System," *Proc. of 5th Sym. on Wavelets, Subband and Block Transforms in Commun.*, NJ, March 1997.

Analysis of a Distributed Power Control Algorithm*

Jeffrey D. Herdtner and Edwin K. P. Chong
School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907-1285
{herdtner, echong}@ecn.purdue.edu

Introduction: We consider a wireless communication system consisting of mobile users communicating to base stations over a common radio channel. We analyze a low complexity uplink power control algorithm. At each iteration, each user needs only a single bit of information from its assigned base station indicating whether its SIR is above or below its desired level. If it is above, the user decreases its power level by a fixed amount (in dB); otherwise, the power level is increased by a fixed amount. This type of algorithm has been shown to be effective in simulation studies of various types of systems [1,3,4]. Because of the fixed step sizes, the algorithm does not actually converge to the optimal power assignment, but we prove its stability and some performance guarantees in terms of analytically derived bounds on the convergence error. We show that the convergence error can be made as small as desirable, at the cost of reducing the convergence rate. We do not consider the effects of mobility, although this has been investigated via simulations in [1].

The Power Control Problem: Consider a wireless system with M base stations and N users communicating over a common radio channel. The assignment of users to base stations is assumed to be fixed. The SIR of user i (at its assigned base station) is

$$SIR_i(p) = \frac{h_{ii}p_i}{\sum_{j\neq i} h_{ji}p_j + \sigma_i},$$

where p_i is the transmit power level of user i, h_{ij} is the gain from user i to the assigned base station of user j, σ_i is the receiver noise at the assigned base station of user i, and $\boldsymbol{p} = [p_1, \cdots, p_N]^T$. Let γ_i be the desired SIR of user i. The power control problem is that of finding the smallest power vector, $\boldsymbol{p}$, such that $SIR_i(\boldsymbol{p}) \geq \gamma_i$, $\forall i$. Letting $\boldsymbol{H}$ be the $N \times N$ matrix with entries $H_{ii} = 0$ and $H_{ij} = \gamma_i h_{ji}/h_{ii}$ for $j \neq i$, and $\boldsymbol{\delta}$ be the N-vector with entries $\delta_i = \gamma_i\sigma_i/h_{ii}$, the power control problem can be written as: minimize $\boldsymbol{p}$ subject to $\boldsymbol{p} \geq \boldsymbol{H}\boldsymbol{p} + \boldsymbol{\delta}$, $\boldsymbol{p} > \boldsymbol{0}$, where the inequalities are componentwise and the minimization is with respect to the partial order associated with $\leq$. (Previous work using this formulation includes [2,5].) It is known that a unique optimal (minimum power) solution, $\boldsymbol{p}^*$, to the power control problem exists if and only if $\rho(\boldsymbol{H}) < 1$, where $\rho(\boldsymbol{H})$ is the spectral radius of $\boldsymbol{H}$.

Algorithm Description: We now describe the algorithm that we will analyze. At each iteration, each user increases its power level by a factor $\beta > 1$ if its SIR is below the

---

*This research supported in part by the National Science Foundation through grant ECS-9501652 and by the U.S. Army Research Office through grant DAAH04-95-1-0246.

desired level; otherwise, the power level is decreased by a factor α, where $0 < \alpha < 1$. To define the algorithm, we first define the vector function $\boldsymbol{T} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ as follows:

$$T_i(\boldsymbol{p}) = \begin{cases} \beta p_i & \text{if } SIR_i(\boldsymbol{p}) < \gamma_i \\ \alpha p_i & \text{if } SIR_i(\boldsymbol{p}) \geq \gamma_i \end{cases}$$

We then let $\boldsymbol{T}(\boldsymbol{p}) = [T_1(\boldsymbol{p}), \ldots, T_N(\boldsymbol{p})]^T$. Given an initial power vector $\boldsymbol{p}^{(0)} > \boldsymbol{0}$, the algorithm generates the sequence of power vectors $\left\{\boldsymbol{p}^{(n)}\right\}_{n=0}^{\infty}$ according to the iteration $\boldsymbol{p}^{(n+1)} = \boldsymbol{T}\left(\boldsymbol{p}^{(n)}\right)$, $n \geq 0$.

This algorithm has very desirable characteristics. No global information is needed for users to update their power levels, and each user needs only a single bit of information per iteration from its assigned base station, indicating whether or not its SIR constraint is satisfied. It is assumed that the SIR is accurately measured and that the corresponding information bit is transmitted without error to the user, but simulations in [1,4] indicate that this type of algorithm is robust in less ideal conditions.

Stability and Convergence Analysis: An algorithm is considered to be stable if, for any initial condition $\boldsymbol{p}^{(0)}$, there is a finite power vector $\overline{\boldsymbol{p}}$ such that $\boldsymbol{0} \leq \boldsymbol{p}^{(n)} \leq \overline{\boldsymbol{p}}$ for all $n \geq 0$. In this work, we prove that the power control algorithm defined above is stable provided $\beta < 1/\rho(\boldsymbol{H})$. The proof proceeds by constructing a set that contains $\boldsymbol{p}^{(0)}$ and is invariant under $\boldsymbol{T}$. For details, contact the authors.

While stability is important, it does not guarantee good performance. However, assuming $\beta < 1/\rho(\boldsymbol{H})$, we prove that the power control algorithm provides users with acceptable connections after a finite number of iterations. Specifically, we show that for each user i, $SIR_i\left(\boldsymbol{p}^{(n)}\right) \geq (\alpha/\beta)\gamma_i$ for n large. By choosing α and β close to one, all users at least come close to meeting their desired SIR.

It is desirable not only to provide acceptable connections for all users, but also to prevent excess power usage. Assuming $\beta < 1/\rho(\boldsymbol{H})$, we establish a bound on the amount the algorithm exceeds the optimal power assignment. Specifically, we show that $\limsup \boldsymbol{p}^{(n)} \leq \overline{\boldsymbol{p}}(\beta)$, where $\overline{\boldsymbol{p}}(\beta) \triangleq \beta(\boldsymbol{I} - \beta\boldsymbol{H})^{-1}\boldsymbol{\delta}$. Since it can easily be shown that $\overline{\boldsymbol{p}}(\beta) \searrow \boldsymbol{p}^*$ as $\beta \searrow 1$, the error is controllable by the choice of β. However, there is a tradeoff in that reducing the error (i.e., decreasing β) results in slower convergence.

References:

[1] S. Ariyavisitakul, "Signal and interference statistics of a CDMA system with feedback power control - Part II," *IEEE Trans. on Comm.*, vol. 42, pp. 597–605, 1994.

[2] S. V. Hanly, "An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity," *IEEE J. on Sel. Areas in Comm.*, vol. 13, no. 7, pp. 1332–1340, Sept. 1995.

[3] F. Ling, B. Love, and M. M. Wang, "Behavior and performance of power controlled IS-95 reverse-link under soft handoff," in *Proc. IEEE Veh. Tech. Conf.*, vol. 2, pp. 924–928, Sept. 1997.

[4] A. J. Viterbi, A. M. Viterbi, and E. Zehavi, "Performance of power-controlled wideband terrestrial digital communication," *IEEE Trans. on Comm.*, vol. 41, no. 4, pp. 559–569, Apr. 1993.

[5] R. D. Yates and C.-Y. Huang, "Integrated power control and base station assignment," *IEEE Trans. on Veh. Tech.*, vol. 44, no. 3, pp. 638–644, Aug. 1995.

Adaptive Multiuser Decision Feedback Demodulation for GSM

Rapeepat Ratasuk and Michael L. Honig
Department of ECE, Northwestern University
2145 Sheridan Road, Evanston, IL 60208

Abstract

Demodulation of multiple users in a cellular TDMA system with large power imbalance among the received signals and limited side information is considered. Our approach is to combine an adaptive multiuser decision-feedback demodulator (DFD) with multiple antennas. The DFD contains (matrix) finite-impulse response feedforward and and feedback filters to mitigate both intersymbol and multiple-access interference. The MMSE filter coefficients are specified, and adaptive estimation algorithms are presented which rely on training data from each user. Simulation results show that by cancelling a relatively strong user, the DFD offers a significant improvement in performance for weak users relative to the linear MMSE receiver.

1 Introduction

We consider adaptive space-time multiuser demodulation of digital signals in a cellular environment with limited side information. The communication system of interest is the forward link of a cellular Time-Division Multiple-Access (TDMA) network (in particular, Global System for Mobile telecommunications (GSM)). The receiver is assumed to be randomly located within a particular cell, and has no prior knowledge of users' spatial signatures or channels. Since the co-channel users to be demodulated are in different cells, there is a significant near-far problem which causes a large power imbalance among the received signals.

The approach we take is to use multiuser decision-feedback demodulation combined with spatial diversity. Decision-feedback equalization for multiple-input/multiple output (MIMO) channels was presented in [1], and applied to multiuser detection for DS-CDMA in [2, 3, 4].

The emphasis of our study is on an *adaptive* multiuser decision-feedback demodulator (DFD), which promises superior performance to an adaptive linear Minimum Mean-Square Error (MMSE) receiver, at the cost of a moderate increase in complexity. The feedforward filter of the adaptive DFD consists of an adaptive *linear* MMSE filter followed by an adaptive whitening filter. For the channel model considered, both multiple-access and intersymbol interference are present. (That is, all users are subject to slow frequency-selective fading.) The optimal infinite-impulse-response (IIR) feedforward and feedback filters [3] are approximated as finite-impulse-response (FIR) filters [5, 6]. Embedding the received vectors in a higher-dimensional space leads to a convenient method for estimating the optimal (matrix) filter coefficients.

The main disadvantage of the DFD is its sensitivity to error propagation. Performance can degrade substantially when the feedback error rate is high. In our scenario, however, error propagation is alleviated via proper ordering of users (strongest to weakest) and successive cancellation. That is, strong users are demodulated first, allowing weaker users to make use of those decisions for interference cancellation.

In the next section, we present the communication system model. In Section 3, we present the MMSE-DFD, along with the embedded representation of the filters for frequency-selective fading channels. This is followed, in Section 4, by a discussion of adaptive Least-Squares (LS) algorithms for estimating the filter coefficients. Simulation results are presented in Section 5, and conclusions are presented in Section 6.

2 System Model

We assume that the receiver is randomly located in a cell and attempts to demodulate the desired user within that cell, in addition to co-channel users in other cells. The other-cell users are likely to be located much further away from the receiver than the adjacent base station, creating a large disparity in received powers. GMSK modulation is assumed, which is a nonlinear technique. However, it has been shown that a GMSK signal can be accurately approximated with the standard linear baseband signal [7, 8]

$$s_k(t) = \sum_i b_k(i) g(t - iT - \tau_k) \tag{1}$$

for user k, where $b_k(i)$ is the i^{th} transmitted bit for user k, $g(t)$ is the transmitted pulse shape, T is the symbol period, and τ_k is the random delay associated with the k^{th} user. In a frequency-selective fading environment, the received signal at the m^{th} antenna is given by

$$r_m(t) = \sum_{k=1}^{K} \sum_{l=1}^{L} h_{m,k}^{(l)}(t) s_k(t - \tau_m^{(l)}) + n_m(t) \tag{2}$$

where $h_{m,k}^{(l)}(t)$ is the complex fading coefficient associated with user k and path l, $n_m(t)$ is additive white complex-Gaussian noise with variance σ^2, $\tau_m^{(l)}$ is the delay associated with path l on antenna m, L is the number of paths, and K is the number of users. We assume independent Rayleigh fading on each antenna element. In the case of flat fading, the spatial signature for a particular user is the vector of complex fade coefficients across the antennas. All channels are assumed to be constant throughout each packet, but are statistically independent from packet to packet.

When the received signal is sampled at the symbol rate, (2) may be expressed in vector form as

$$\mathbf{r}(i) = \sum_{l=1}^{L_c} \mathbf{H}^{(l)} \mathbf{b}(i - l + 1) + \mathbf{n}(i) \tag{3}$$

where the k^{th} column of the matrix

$$\mathbf{H}^{(l)} = \begin{bmatrix} h_{1,1}^{(l)} & h_{1,2}^{(l)} & \cdots & h_{1,K}^{(l)} \\ h_{2,1}^{(l)} & h_{2,2}^{(l)} & \cdots & h_{1,K}^{(l)} \\ \vdots & \vdots & \ddots & \vdots \\ h_{M,1}^{(l)} & h_{M,2}^{(l)} & \cdots & h_{M,K}^{(l)} \end{bmatrix} \tag{4}$$

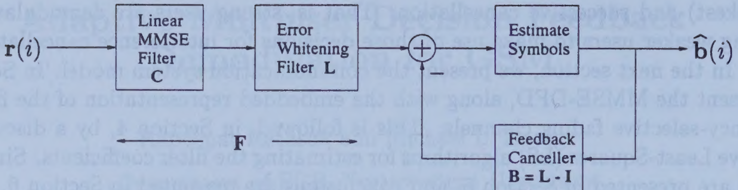


Figure 1: Decision-Feedback Demodulator Block Diagram

is the spatial signature for user k, path l, and L_c represents the overall channel order. In this case, $\mathbf{H}^{(l)}$ incorporates both the channel response and the intersymbol interference due to the Gaussian pulse shaping filters.

3 Multiuser MMSE-DFD

A block diagram of the multiuser DFD is shown in Figure 1. The soft output at time i is

$$\mathbf{x}(i) = \mathbf{F}\mathbf{r}(i) - \mathbf{B}\widehat{\mathbf{b}}(i) \tag{5}$$

where $\mathbf{F}$ and $\mathbf{B}$ are respectively the feedforward and feedback filters, $\mathbf{B}$ is lower diagonal (with zeros along the diagonal), and $\widehat{\mathbf{b}}(i)$ is the output of the decision device. The performance criterion used to optimize the receiver is Minimum Mean Squared Error (MMSE), defined as

$$\min_{\mathbf{F},\mathbf{B}} \mathcal{E} = \min_{\mathbf{F},\mathbf{B}} E\left[\|\mathbf{b}(i) - \mathbf{x}(i)\|^2\right] \tag{6}$$

For tractability, it is typically assumed that $\widehat{\mathbf{b}}(i) = \mathbf{b}(i)$ (i.e., no error propagation), in which case the optimal filters are [1]

$$\mathbf{F} = (\mathbf{I} + \mathbf{B})\mathbf{P}^{\dagger}\mathbf{R}^{-1} \tag{7}$$

where

$$\mathbf{P} = E[\mathbf{r}(i)\mathbf{b}^{\dagger}(i)], \qquad \mathbf{R} = E[\mathbf{r}(i)\mathbf{r}^{\dagger}(i)] \tag{8}$$

and

$$\mathbf{B} = \mathbf{L} - \mathbf{I} \tag{9}$$

where $\mathbf{L}$ is a lower-diagonal matrix which satisfies

$$\mathbf{L}^{\dagger}\mathbf{D}\mathbf{L} = \mathbf{P}^{\dagger}\mathbf{P} + \sigma^2\mathbf{I} \tag{10}$$

where $\mathbf{D}$ is diagonal, and the matrix in (10) is the error covariance matrix associated with the linear MMSE receiver. The users are demodulated successively as

$$\begin{aligned}\hat{b}_1(i) &= \operatorname{sgn}\left[\operatorname{Re}\left\{z_1(i)\right\}\right]\\ \hat{b}_2(i) &= \operatorname{sgn}\left[\operatorname{Re}\left\{z_2(i) - \mathbf{B}_{2,1}\hat{b}_1(i)\right\}\right]\\ &\vdots\\ \hat{b}_K(i) &= \operatorname{sgn}\left[\operatorname{Re}\left\{z_K(i) - \sum_{k=1}^{K-1}\mathbf{B}_{K,K-k}\hat{b}_k(i)\right\}\right]\end{aligned} \tag{11}$$

where

$$z_k(i) = \mathbf{L}\mathbf{C}_k^{\dagger}\mathbf{r}(i) \tag{12}$$

is the output of the feedforward filter for user k.

Multipath introduces interference across time in addition to interference across users. The ideal MMSE-DFD in this case has matrix IIR feedforward and feedback filters with transfer functions $\mathbf{F}(z)$ and $\mathbf{B}(z)$ [1]. Here we instead assume Finite-Impulse Response (FIR) filters, which we optimize according to the MMSE criterion. To do this we define the stacked symbol vector

$$\widetilde{\mathbf{b}}(i) = \left[b_1(i - L_c + 1) \ \cdots \ b_K(i - L_c + 1) \ \cdots \ b_1(i) \ \cdots \ b_K(i)\right]^T \tag{13}$$

where the users are ordered according to received power, and across time. In what follows, all variables with a tilde are stacked variables. The intersymbol interference can be treated as interference from separate, virtual users. The cost function then becomes

$$\min_{\widetilde{\mathbf{F}},\widetilde{\mathbf{B}}} \mathcal{E} = \min_{\widetilde{\mathbf{F}},\widetilde{\mathbf{B}}} E\left(\left\|\mathbf{b}(i) - \left(\widetilde{\mathbf{F}}\widetilde{\mathbf{r}}(i) - \widetilde{\mathbf{B}}\widetilde{\mathbf{b}}(i)\right)\right\|^2\right) \tag{14}$$

where

$$\widetilde{\mathbf{F}} = \left[\mathbf{F}_{L_c-1} \ \ \mathbf{F}_{L_c-2} \cdots \ \ \mathbf{F}_0\right] \tag{15}$$

$$\widetilde{\mathbf{B}} = \left[\mathbf{B}_{L_c-1} \ \ \mathbf{B}_{L_c-2} \ \ \ldots \ \ \mathbf{B}_0\right] \tag{16}$$

and $\mathbf{F}_n$ and $\mathbf{B}_n$ represent the feedforward and feedback filters at sampling instance n, respectively. The stacked received vector is given by

$$\widetilde{\mathbf{r}}(i) = \widetilde{\mathbf{H}}\widetilde{\mathbf{b}}(i) + \widetilde{\mathbf{n}}(i) \tag{17}$$

where $\widetilde{\mathbf{H}}$ represent the overall response of the communication channel.

As before, the solutions to the higher dimensional minimization problem are

$$\mathcal{F} = (\mathbf{I} + \mathbf{B})\mathbf{P}^{\dagger}(\widetilde{\mathbf{H}}^{\dagger}\widetilde{\mathbf{H}} + \sigma^2\mathbf{I})^{-1}, \qquad \mathcal{B} = \mathbf{L} - \mathbf{I} \tag{18}$$

where

$$\mathbf{L}^{\dagger}\mathbf{D}\mathbf{L} = (\widetilde{\mathbf{H}}^{\dagger}\widetilde{\mathbf{H}} + \sigma^2\mathbf{I}) \tag{19}$$

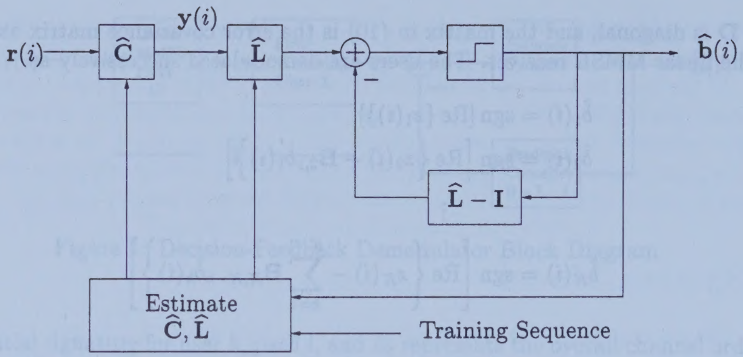

$\mathbf{y}(i)$
$\mathbf{r}(i)$
$\widehat{\mathbf{C}}$
$\widehat{\mathbf{L}}$
$-$
$\widehat{\mathbf{b}}(i)$
$\widehat{\mathbf{L}}-\mathbf{I}$
Estimate
$\widehat{\mathbf{C}}, \widehat{\mathbf{L}}$
Training Sequence

Figure 2: Least-Squares DFD

The solutions to the original (lower-dimensional) problem are then determined as follows:

$$\mathcal{B} = \begin{bmatrix} \mathcal{B}_1 & \mathbf{0} & \mathbf{0} & \dots \\ \mathcal{B}_2 & \mathcal{B}_3 & \mathbf{0} & \dots \\ & \vdots & \ddots & \\ \mathbf{B}_{L_c-1} & \mathbf{B}_{L_c-2} & \dots & \mathbf{B}_0 \end{bmatrix} \} \widetilde{\mathbf{B}} \qquad \mathcal{F} = \begin{bmatrix} \mathcal{F}_1 & \mathbf{0} & \mathbf{0} & \dots \\ \mathcal{F}_2 & \mathcal{F}_3 & \mathbf{0} & \dots \\ & \vdots & \ddots & \\ \mathbf{F}_{L_c-1} & \mathbf{F}_{L_c-2} & \dots & \mathbf{F}_0 \end{bmatrix} \} \widetilde{\mathbf{F}} \tag{20}$$

The users are then demodulated successively as

$$\begin{aligned} \hat{b}_1(i) &= \operatorname{sgn}\left[z_1(i) - \sum_{j=0}^{L_c-1} \mathbf{B}_j(1,1)\hat{b}_1(i-j) \right] \\ &\vdots \\ \hat{b}_K(i) &= \operatorname{sgn}\left[z_K(i) - \sum_{j=0}^{L_c-1} \sum_{k=1}^{K} \mathbf{B}_j(K,k)\hat{b}_k(i-j) \right] \end{aligned} \tag{21}$$

In this way, interference from stronger users generated at times $i - L_c + 1, \dots, i$ is cancelled. Note, however, that the dimension of the signal space has increased, and hence more data are required to estimate $\widetilde{\mathbf{H}}$. Also, errors propagate in time and across users.

4 Multiuser LS-DFD

In channel parameters are unknown, we present several adaptive algorithms for estimating the filter coefficients, shown in Figure 2. In this case, a training sequence available in GSM packets is used to estimate the filter coefficients. Once estimates of these coefficients are available, the DFD can be used to obtain initial bit estimates. These bit estimates may then be as the input to a decision-directed least squares (LS) algorithm, which provides refined bit estimates for the packet of interest. This process can be iterated until the bit estimates remain fixed.

In this section we present three variants of LS algorithms that can be used to estimate the MMSE-DFD coefficients.

Decision-Directed LS

We wish to minimize the LS cost function, i.e.,

$$\min_{\widehat{\mathbf{F}},\widehat{\mathbf{B}}} \mathcal{E} = \min_{\widehat{\mathbf{F}},\widehat{\mathbf{B}}} \sum_{i=1}^{B} \|\mathbf{b}(i) - (\widehat{\mathbf{F}}\mathbf{r}(i) - \widehat{\mathbf{B}}\widehat{\mathbf{b}}(i))\|^2 \tag{22}$$

where B is the number of bits in a packet. Minimization of this cost function requires either a training sequence or estimates of the transmitted symbols $\mathbf{b}(i)$ from each user. Assuming $\widehat{\mathbf{b}} = \mathbf{b}$, it is easily shown that

$$\widehat{\mathbf{F}} = (\mathbf{I} + \widehat{\mathbf{B}})\widehat{\mathbf{P}}^{\dagger}\widehat{\mathbf{R}}^{-1}, \tag{23}$$

where

$$\widehat{\mathbf{P}} = \frac{1}{B}\sum_{i=1}^{B} \mathbf{r}(i)\mathbf{b}^{\dagger}(i), \qquad \widehat{\mathbf{R}} = \frac{1}{B}\sum_{i=1}^{B} \mathbf{r}(i)\mathbf{r}^{\dagger}(i) \tag{24}$$

Defining the estimation error

$$\zeta(i) = \mathbf{b}(i) - \widehat{\mathbf{P}}^{\dagger}\widehat{\mathbf{R}}^{-1}\mathbf{r}(i), \tag{25}$$

the sample error covariance at the output of the feedforward filter is given by

$$\mathcal{E} = \mathbf{L}\left(\frac{1}{B}\sum_{i=1}^{B} \zeta(i)\zeta^{\dagger}(i)\right)\mathbf{L}^{\dagger} \tag{26}$$

where the lower-diagonal matrix

$$\widehat{\mathbf{L}} = \mathbf{I} + \widehat{\mathbf{B}}. \tag{27}$$

To minimize the LS cost function, we choose $\mathbf{L}$ to whiten the sample error covariance matrix,

$$\widehat{\mathbf{L}}^{-\dagger}\widehat{\mathbf{D}}\widehat{\mathbf{L}}^{-1} = \frac{1}{B}\sum_{i=1}^{B} \zeta(i)\zeta^{\dagger}(i) \tag{28}$$

where $\mathbf{D}$ is diagonal. A decision-directed version of the preceding algorithm is obtained by using estimates from the linear LS filter:

$$\widehat{\mathbf{b}}(i) = \operatorname{sgn}\left[\operatorname{Re}\left\{\widehat{\mathbf{P}}^{\dagger}\widehat{\mathbf{R}}^{-1}\mathbf{r}(i)\right\}\right] \tag{29}$$

Channel Estimation

An alternative to the preceding LS approach is to estimate the channel parameters associated with the users, and then use these to estimate the MMSE filters. (This approach is taken in [4]). In this case, we select $\mathbf{L}$ to satisfy

$$\widehat{\mathbf{L}}^{-1}\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{L}}^{-\dagger} = (\widehat{\mathbf{P}}^{\dagger}\widehat{\mathbf{P}} + \widehat{\sigma}^2\mathbf{I})^{-1} \tag{30}$$

where $\widehat{\mathbf{P}}$ is given by (24), and $\hat{\sigma}$ is an estimate of the noise variance σ. The matrix in (30) is an estimate of the error covariance matrix in (10). The noise variance can be estimated as

$$\hat{\sigma}^2 = \frac{1}{K}\,\mathrm{tr}(\widehat{\mathbf{R}} - \widehat{\mathbf{P}}\widehat{\mathbf{P}}^\dagger) \tag{31}$$

Blind Estimation

In the absence of training bits or reliable symbol estimates, it is possible to compute a blind estimate of the whitening filter. Namely, expanding the error covariance matrix gives

$$\begin{aligned}\frac{1}{B}\sum_{i=1}^{B}\zeta(i)\zeta^\dagger(i) = \frac{1}{B}\sum_{i=1}^{B}\Big[&\widehat{\mathbf{b}}(i)\widehat{\mathbf{b}}^\dagger(i) - \widehat{\mathbf{b}}(i)\mathbf{r}^\dagger(i)\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{P}} - \\ &\widehat{\mathbf{P}}^\dagger\widehat{\mathbf{R}}^{-1}\mathbf{r}(i)\widehat{\mathbf{b}}^\dagger(i) + \widehat{\mathbf{P}}^\dagger\widehat{\mathbf{R}}^{-1}\mathbf{r}(i)\mathbf{r}^\dagger(i)\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{P}}\Big]\end{aligned} \tag{32}$$

Since the bits are assumed to be i.i.d, the first term in (32) may be approximated by

$$\frac{1}{B}\sum_{i=1}^{B}\widehat{\mathbf{b}}(i)\widehat{\mathbf{b}}^\dagger(i) \approx \mathbf{I}. \tag{33}$$

Substituting this approximation and the estimate for $\widehat{\mathbf{P}}$ in (24) into (32) gives

$$\frac{1}{B}\sum_{i=1}^{B}\zeta(i)\zeta^\dagger(i) \approx \mathbf{I} - \frac{1}{B}\sum_{i=1}^{B}\mathbf{y}(i)\mathbf{y}^\dagger(i) \tag{34}$$

where

$$\mathbf{y}(i) = \widehat{\mathbf{P}}^\dagger\widehat{\mathbf{R}}^{-1}\mathbf{r}(i) \tag{35}$$

In a stationary environment this estimate converges to the error covariance matrix as $B \to \infty$. The matrix $\widehat{\mathbf{L}}$ is then selected to satisfy

$$\widehat{\mathbf{L}}^{-\dagger}\widehat{\mathbf{D}}\widehat{\mathbf{L}}^{-1} = \mathbf{I} - \frac{1}{B}\sum_{i=1}^{B}\mathbf{y}(i)\mathbf{y}^\dagger(i) \tag{36}$$

To summarize, we have presented three different estimators for the MMSE-DFD. All converge to the MMSE solution as the number of data samples increases to infinity. A comparison of these techniques is given in Section 5.

5 Numerical Results

The results in this section assume a fixed number of users, where one user is relatively strong (inside the cell), and the rest are weak (outside the cell). The modulation scheme is GMSK, where we use the linear model presented in [7]. The time slot is a normal burst, as defined by the GSM standard, and the receiver is synchronous with respect to the users. Since the linearized GMSK pulse shaping filter is partial response, $\mathbf{r}(i)$ contains both intersymbol and multiple-access interference. All users experience a frequency-selective Rayleigh fading channel with three rays one symbol interval apart. The fading coefficients associated with the paths and antennas are assumed to be independent, and are constant for the duration of the packet.

Figure 3 shows uncoded error rate vs. SNR for the MMSE-DFD and linear receivers. In this case, there are two users, two antennas, and the observation window spans 5 symbols. The strong user is 10 dB stronger than the weak user. (That is, the SNR for user 1 shown on the x-axis is $\mathrm{SNR}_1 = \mathrm{SNR}_2 + 10\mathrm{dB}$). The figure shows that the MMSE-DFD offers a significant improvement for the weak user, relative to linear detection, since the MMSE-DFD cancels interference from the strong user, in addition to ISI. The strong user also exhibits some improvement due to cancellation of ISI.

Figure 4 illustrates the performance of the MMSE-DFD in a rank-deficient system with two antennas and three users. The two weaker users are received with the same SNR, while the strongest user is 10 dB stronger than the weak users. The two weak users have identical performance with the linear MMSE receiver. Again, the MMSE-DFD offers a significant performance improvement for the weak users, since the strong user is effectively cancelled. Note that the performance of the last user is better than that of the second user since more interferers are cancelled.

Finally, Figure 5 shows error rate as a function of training data for the three adaptive algorithms discussed in Section 4. In this case, the decision-directed LS-DFD converges relatively quickly compared with the other techniques.

6 Conclusions

We have studied the application of the MMSE-DFD receiver with multiple antennas to multiuser demodulation of GMSK signals. It has been shown that for the cases considered, where there is a large power imbalance among received users, the MMSE-DFD offers a significant performance improvement for the weak users relative to the MMSE linear receiver. Adaptive estimation algorithms have been presented with complexity comparable to that of analogous linear LS algorithms. (An additional whitening operation is required for the DFD.) The filter structure considered includes memory, which is needed to mitigate the effects of multipath-induced ISI.

The main disadvantage of the DFD relative to a linear detector is error propagation. Results show that when the SNR for users being cancelled is low (say, less than 5 dB), error propagation is likely to increase the error rates of successive users relative to linear detection.

As the number of DFD coefficients increases (due to both ISI and more users), more training data is needed to obtain an accurate estimate of the MMSE coefficients. This is a potential problem for the application considered, in which relatively little training data is available. Future efforts will include consideration of other adaptive techniques, and a comparison of the MMSE-DFD with maximum-likelihood detection.

References

[1] A. Duel-Hallen, "Equalizers for Multiple Input/Multiple Output Channels and PAM Systems with Cyclostationary Input Sequences," *IEEE Journal on Selected Areas in Communications*, vol. 10, pp. 630–639, Apr. 1992.

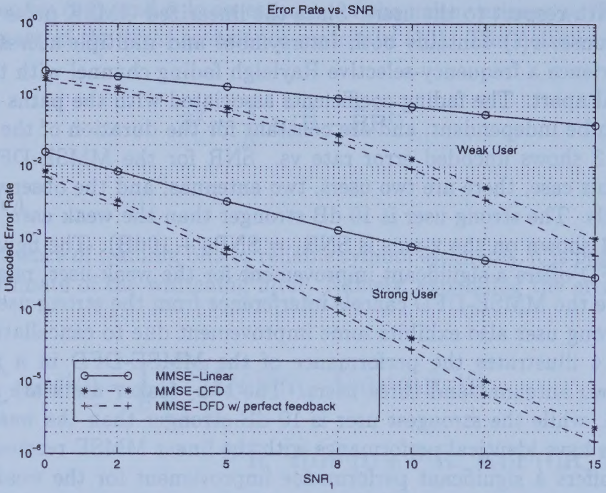


Figure 3: Uncoded error rate for a 2-user, 3-ray multipath GSM channel with 2 antenna elements.

[2] A. Duel-Hallen, "Decorrelating Decision-Feedback Multiuser Detector for Synchronous Code-Division Multiple Access Channel," *IEEE Transactions on Communications*, vol. 41, pp. 285–290, Feb. 1993.

[3] A. Duel-Hallen, "A Family of Multiuser Decision-Feedback Detectors for Asynchronous Code-Division Multiple-Access Channels," *IEEE Transactions on Communications*, vol. 43, pp. 421–434, feb/mar/apr 1995.

[4] A. Klein, G. K. Kaleh, and P. W. Baier, "Zero Forcing and Minimum Mean-Square-Error Equalization for Multiuser Detection in Code-Division Multiple-Access," *IEEE Transactions on Vehicular Technology*, vol. 45, pp. 276–287, May 1996.

[5] C. Belfoire and J. Park, Jr., "Decision Feedback Equlization," *Proceedings of the IEEE*, vol. 67, pp. 1143–1156, Aug. 1979.

[6] N. Al-Dhahir and J. M. Cioffi, "MMSE Decision-Feedback Equalizers: Finite-Length Results," *IEEE Transactions on Information Theory*, vol. 41, pp. 961–975, July 1995.

[7] J. W. Liang, B. C. Ng, J. T. Chen, and A. Paulraj, "GMSK Linearization and Structured Channel Estimate for GSM Signals," *MILCOM 97 Proceedings*, pp. 817–821, Nov. 1997.

[8] G. Castellini, F. Conti, E. Del Re, and L. Peirucci, "A Continously Adaptive MLSE Receiver for Mobile Communications: Algorithm and Performance," *IEEE Transactions on Communications*, vol. 45, pp. 80–89, Jan. 1997.

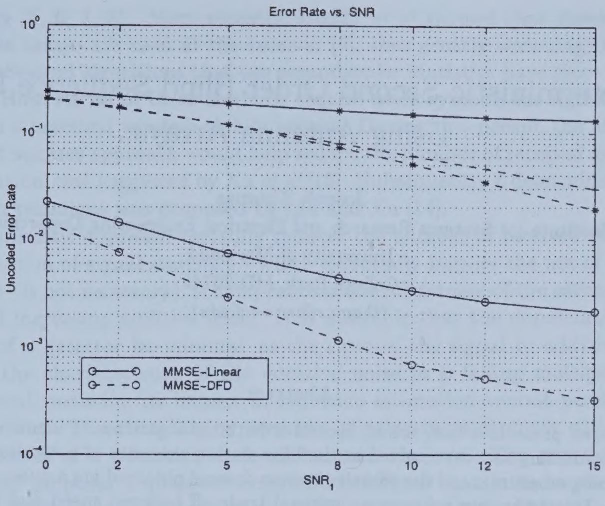


Figure 4: Uncoded error rate for a rank-deficient system: 3 users, 2 antenna elements, 3-ray multipath GSM channel. $\mathrm{M}_k, \mathrm{D}_k$ = Performance of user k under linear MMSE and MMSE-DFD receivers, respectively.

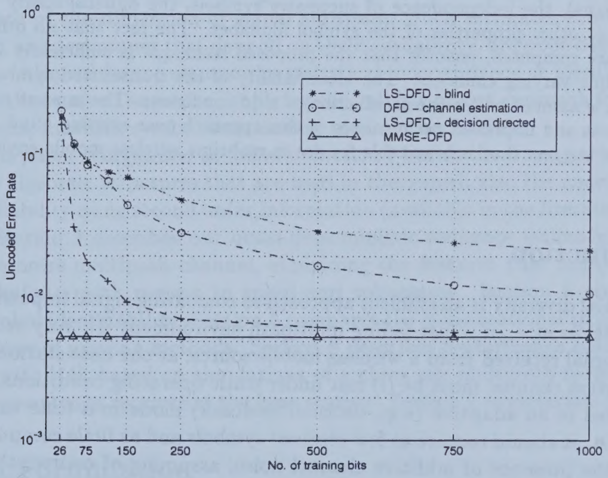


Figure 5: Uncoded error rate convergence for various adaptive DFD algorithms, 2 users (weak user shown), 3-ray multipath GSM channel, 2 antenna elements, $\mathrm{SNR}_2 = 2$ dB.

Quasi-Deterministic Second Order Blind Sequence Estimation Using Antenna Arrays

Joseph Thomas
Institute for Systems Research *and* Electrical Engineering Department
University of Maryland
College Park, MD 20742
(thomas@isr.umd.edu)

Abstract

This paper presents a fast, robust second order blind algorithm to identify the symbols from a transmitting source, received at the base station antennas of a wireless reverse link channel, using an estimate of the effective inverse channel obtained via a quasi-deterministic approach. This technique achieves an optimal trade-off between speed and robustness in the presence of increasing additive channel noise by using estimates of the additive noise variance and short term temporally averaged estimates of an analogue of the ensemble averaged autocorrelation of the received sequence. Using these psuedo-autocorrelations provides a faster sequence estimation procedure than existing deterministic techniques which merely seek to use the properties of the signal subspace, ignoring the correlation between successive received symbols - an approach that has naturally been confined to statistical techniques. Likewise, including the additive noise variance (which needs to be computed rather infrequently compared with the block processing rate) lends robustness to the sequence estimation procedure. No assumptions are made however, regarding the distribution of the received signal, the independence of successive symbols, the dimensionality of the signal subspace, or other properties of the symbol alphabet. The fact that no other statistical averages are computed, ensures that the proposed technique is sufficiently fast for operation in time varying channels. The identifiability of the transmitted symbols using this technique, is assured under the satisfaction of mild conditions. The modest computational requirements and improved performance of this approach over existing blind identification techniques, suggest that it is suitable for use in real-time wireless mobile environments.

1 Introduction

Exploiting spatial diversity is an efficient approach to combating signal impairments arising from intersymbol interference, in a fast fading multipath environment [1]. Any technique that seeks to detect the signal received from a wireless mobile source, at the base station receiver, through such a propagation channel must be (i) fast under static operating conditions, so that it may be usefully operated in an adaptive (e.g., decision feedback) mode in a time varying propagation environment, i.e., it should require as few received symbols and as little computation as possible, (ii) robust in the presence of additive channel noise, assuming of course, that the estimation process satisfies the zero-forcing property, and (iii) bandwidth efficient, so that the overheads which arise from the need for transmitting training sequences are minimised, effectively implying blind or semi-blind estimation.

The requirement of speed eliminates the possibility of considering bussgang [2] or cumulant-based techniques [3]. Following seminal work by Gardner [4] and Tong *et al* [5], the use of second order statistics (which require fewer symbols for their estimation than do cumulants) for blind identification of communication channels was understood, as were the conditions for such identifiability [5, 6, 7, 8]. More recently, Reddy *et al* showed that these conditions are satisfied if antenna arrays are used at the receiver [9], thus greatly widening the scope of this class of methods. Several algorithms that use second-order statistics have since been proposed, e.g., [10, 11, 12]. However all of these methods require more symbols for signal detection than can be afforded in a practical wireless mobile system. During this period, the idea that using a deterministic least squares approach would help reduce the number of received symbols required for channel estimation, was suggested by Xu *et al* [13]. Subsequently a deterministic direct blind symbol estimation technique was proposed by Liu and Xu [14].

While a deterministic approach certainly reduces the number of received symbols required (since the computation of signal statistics - which essentially implies the use of a large number of received signals - is not necessary), it compromises the robustness of the estimation technique in the presence of increasing additive noise. It is desirable that the departure from the zero-forcing property of a detector be minimal, as the ratio of the signal to additive noise powers increases. Since the characteristics of the additive noise in a fading multipath channel are essentially statistical, reducing its impact in sequence estimation cannot possibly be effected in a deterministic way. Therefore this requirement of robustness conflicts with that of speed, implying the existence of a potential trade-off between the two. The present work explores this possibility, and proposes a quasi-deterministic technique that includes the noise variance in its formulation, without requiring the computation of the statistics of the signal. In contrast with existing deterministic channel or sequence estimation techniques that seek to exploit only the properties of the signal subspace from the input-output relation of the system model, the present work introduces in its formulation, the correlation between successive symbols in the form of a short term temporally averaged analogue of the ensemble averaged autocorrelation function. The resulting method is both fast and adequately robust at low signal to additive noise ratios (SANRs), in contrast with the sequence estimation method in [14]. Besides providing significantly better performance over [14], the computational requirements of our approach are much lower. In contrast with many existing channel estimation techniques [5, 10], we do not require an estimate of the dimension of signal subspace. Such estimates are very difficult to obtain in practice, and techniques that depend on their availability, are poor candidates for real-time applications. Also we do not assume the statistical independence or identical distribution of the symbols in the received sequence, implying the validity of our approach for certain coded modulation systems, in which symbol interleaving may not be feasible.

The rest of this paper is organised as follows. Section 2 builds the framework of our formulation, defining the algebraic structures that are used in the sequel; also the necessary conditions for symbol identifiability using second order information (available in the literature) are adapted to our setting. Section 3 describes our quasi-deterministic sequence estimation approach for an additive white noise multipath channel, explaining the features that imply its superiority over existing techniques with respect to speed and robustness. Section 4 provides sufficient conditions for symbol identifiability. Section 5 compares the computational requirements of our technique with those of existing methods. Section 6 describes performance studies of our technique compared with [14], based on simulations of a three ray multipath reverse link wireless channel with a four-antenna receiver at the base station. Section 7 summarises our conclusions.

2 Problem Formulation

A wireless reverse link channel with a single transmitting source is often modelled as a single-input multiple-output (SIMO) system. The channel consists of the modulator, the transmit filters, the physical propagation path, the receive filters, and the matched-filter demodulators; the component subchannels arise from multiple antenna receivers at the base station. With R receiving antennas, the single-input R-output model includes R parallel subchannels with effective impulse responses described by ${h_n}^{(r)}, 0 \leq r \leq R-1$, and of maximum duration L.

Denoting the transmitted symbol sequence by s_n and the additive noise in the rth subchannel by v_n, the output of the matched filter at the rth antenna element is described by

$$x_n{}^{(r)} = h_n{}^{(r)} \otimes s_n + v_n{}^{(r)} \tag{1}$$

Choosing data blocks of length N, this may be expressed as $\mathbf{x}_n = \mathbf{H}\mathbf{s}_n + \mathbf{v}_n$, or using block hankel matrices, constructed from M such blocks, as

$$\mathbf{X}_n = \mathbf{H}\mathbf{S}_n + \mathbf{V}_n \tag{2}$$

where

$$\mathbf{x}_n = \begin{pmatrix} x_n^{(0)} & x_n^{(1)} & \dots & x_n^{(R-1)} & x_{n-1}^{(0)} & x_{n-1}^{(1)} & \dots & x_{n-1}^{(R-1)} & \dots & x_{n-N+1}^{(0)} & x_{n-N+1}^{(1)} & \dots & x_{n-N+1}^{(R-1)} \end{pmatrix}^T,$$

$$\mathbf{s}_n = \begin{pmatrix} s_n & s_{n-1} & \dots & s_{n-N-L+2} \end{pmatrix}^T,$$

$$\mathbf{v}_n = \begin{pmatrix} v_n^{(0)} & v_n^{(1)} & \dots & v_n^{(R-1)} & v_{n-1}^{(0)} & v_{n-1}^{(1)} & \dots & v_{n-1}^{(R-1)} & \dots & v_{n-N+1}^{(0)} & v_{n-N+1}^{(1)} & \dots & v_{n-N+1}^{(R-1)} \end{pmatrix}^T,$$

$$\mathbf{X}_n = \begin{pmatrix} \mathbf{x}_n & \mathbf{x}_{n-1} & \dots & \mathbf{x}_{n-M+1} \end{pmatrix},$$

$$\mathbf{S}_n = \begin{pmatrix} \mathbf{s}_n & \mathbf{s}_{n-1} & \dots & \mathbf{s}_{n-M+1} \end{pmatrix},$$

$$\mathbf{V}_n = \begin{pmatrix} \mathbf{v}_n & \mathbf{v}_{n-1} & \dots & \mathbf{v}_{n-M+1} \end{pmatrix},$$

$$\mathbf{h}_l = \begin{pmatrix} h_l^{(0)} & h_l^{(1)} & \dots & h_l^{(R-1)} \end{pmatrix}^T, 0 \le l \le L-1, \text{ and}$$

$$\mathbf{H} = \begin{pmatrix} \mathbf{h}_0 & \mathbf{h}_1 & \dots & \mathbf{h}_{L-1} & 0 & \dots & \dots & 0 \\ 0 & \mathbf{h}_0 & \mathbf{h}_1 & \dots & \mathbf{h}_{L-1} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \mathbf{h}_0 & \mathbf{h}_1 & \dots & \mathbf{h}_{L-1} \end{pmatrix}$$

Here $\mathbf{H}$ is a block sylvester matrix of dimension $NR \times (N + L - 1)$, and $\mathbf{S}_n$ spans the signal subspace.

Our approach essentially seeks to estimate the received sequence by identifying an effective channel inverse matrix $\mathbf{G}$, which yields the signal subspace matrix $\mathbf{S}_n$, as

$$\mathbf{S}_n = \mathbf{G}^H(\mathbf{X}_n - \mathbf{V}_n) \tag{3}$$

where $\mathbf{G}$ is of dimension $NR \times (N+L-1)$. In the context of traditional approaches, which first estimate the channel and then equalise the system, it has been shown [6, 7] that the channel response is uniquely identifiable from the second order statistics of the received signal under certain conditions. These conditions are necessary for the existence of $\mathbf{G}$, and we therefore have the theorem stated below.

Theorem 1: The effective inverse channel response matrix $\mathbf{G}$ of a single-input R-output system is uniquely identifiable only if the transfer functions, $H^{(r)}(z), 0 \le r \le R-1$, of the component subchannels have no common zeros, or equivalently, the block sylvester channel response matrix $\mathbf{H}$ has full column rank.

A proof of this theorem [6, 7, 8], only assumes uncorrelated signals; in particular the received sequence is not required to be independent or identically distributed - a generality that is retained in our technique, as noted earlier. A sufficient condition for the existence of $\mathbf{G}$ will be developed in Section 4, after a method to obtain it using the received sequence, is described in the next section.

3 Sequence Estimation

Assuming that the effective channel inverse matrix $\mathbf{G}$, exists, we develop a scheme to compute it from the received sequence matrix $\mathbf{X}_n$. Expressing $\mathbf{G}$ as $\begin{pmatrix} \mathbf{g}_0 & \mathbf{g}_1 & \dots & \mathbf{g}_{N+L-2} \end{pmatrix}$ where each $\mathbf{g}_l, 0 \leq l \leq N+L-2$, is a column vector of length NR, we use the hankel property of the signal subspace matrix $\mathbf{S}_n$ (with $M = N+L-1$), in $\mathbf{G}$) to equate the elements on the cross diagonal of the resulting matrix on the right-hand side of (3), and get a chain of equalities

$$\begin{aligned} \mathbf{g}_0^H(\mathbf{x}_{n-M+1} - \mathbf{v}_{n-M+1}) &= \mathbf{g}_1^H(\mathbf{x}_{n-M+2} - \mathbf{v}_{n-M+2}) \\ &= \vdots \\ &= \mathbf{g}_l^H(\mathbf{x}_{n-M+l+1} - \mathbf{v}_{n-M+l+1}) \\ &= \vdots \\ &= \mathbf{g}_{M-1}^H(\mathbf{x}_n - \mathbf{v}_n) \end{aligned} \tag{4}$$

Observe now, that ensemble averaging the outer product of $(\mathbf{x}_{n-k} - \mathbf{v}_{n-k})$ with $(\mathbf{x}_{n-l} - \mathbf{v}_{n-l}), 0 \leq k, l \leq M-1$, with the usual assumption that the signal and noise (which is zero-mean, white) are uncorrelated, yields

$$\begin{aligned} E((\mathbf{x}_{n-k} - \mathbf{v}_{n-k})(\mathbf{x}_{n-l} - \mathbf{v}_{n-l})^H) &= E(\mathbf{x}_{n-k}\mathbf{x}_{n-l}^H) + E(\mathbf{v}_{n-k}\mathbf{v}_{n-l}^H) \\ &= \mathbf{R}_{n-k,n-l}^{(x)} + \sigma_v^2 \mathbf{J}^{l-k} \end{aligned} \tag{5}$$

where $\mathbf{R}_{n-k,n-l}^{(x)}$ is the $NR \times NR$ autocorrelation matrix of the received sequence, σ_v^2 is the variance of the additive noise, and $\mathbf{J}$ is the $NR \times NR$ forward block shifting matrix. We assume $l \geq k$ in the above equation without loss of generality, since the case $l < k$ simply requires taking hermitian transposes. From (4) and (5) we then get

$$\begin{aligned} \mathbf{g}_0^H(\mathbf{R}_{n-M+1,n-r}^{(x)} + \sigma_v^2 \mathbf{J}^{r-M+1}) &= \mathbf{g}_1^H(\mathbf{R}_{n-M+2,n-r}^{(x)} + \sigma_v^2 \mathbf{J}^{r-M+2}) \\ &= \vdots \\ &= \mathbf{g}_l^H(\mathbf{R}_{n-M+l+1,n-r}^{(x)} + \sigma_v^2 \mathbf{J}^{r-M+l+1}) \\ &= \vdots \\ &= \mathbf{g}_{M-1}^H(\mathbf{R}_{n,n-r}^{(x)} + \sigma_v^2 \mathbf{J}^{r}) \\ & \quad 0 \leq r \leq M-1, M = N+L-1 \end{aligned} \tag{6}$$

To extract the information on the inverse channel response, contained in (6), we recast it as

$$\begin{aligned} \mathbf{g}_r(\mathbf{R}_{n-M+r+1,n-M+r+1}^{(x)} + \sigma_v^2 \mathbf{I}) &= \tfrac{1}{M-1} \textstyle\sum_{0 \leq l \leq M-1, l \neq r} \mathbf{g}_l(\mathbf{R}_{n-M+l+1,n-M+r+1}^{(x)} + \sigma_v^2 \mathbf{J}^{r-l}), \\ & \quad 0 \leq r \leq M-1 \end{aligned} \tag{7}$$

These equations may be conveniently organised into the matrix form

$$\mathbf{A}\mathbf{g} = 0 \tag{8}$$

Here $\mathbf{g}$ is a column vector of length $NR(M-1)$ and $\mathbf{A}$ is an $NR(M-1) \times NR(M-1)$ matrix constituted of an $(M-1) \times (M-1)$ block array of $NR \times NR$ autocorrelation matrices, i.e.,

$$\begin{aligned} \mathbf{g} &= \begin{pmatrix} \mathbf{g}_0^H & \mathbf{g}_1^H & \cdots & \mathbf{g}_{M-1}^H \end{pmatrix}^H \\ (\mathbf{A})_{kl} &= \begin{cases} (M-1)(\mathbf{R}_{n-M+k,n-M+k}^{(x)} + \sigma_v^2 \mathbf{I}), & k = l \\ -(\mathbf{R}_{n-M+l,n-M+k}^{(x)} + \sigma_v^2 \mathbf{J}^{k-l}), & k \neq l \end{cases} \end{aligned} \tag{9}$$

so that $\mathbf{A} = \mathbf{Q} + \sigma_v^2 \mathbf{U}$, with

$$(\mathbf{Q})_{kl} = \begin{cases} (M-1)\mathbf{R}_{n-M+k,n-M+k}^{(x)}, & k = l \\ -\mathbf{R}_{n-M+l,n-M+k}^{(x)}, & k \neq l \end{cases}$$

and

$$(\mathbf{U})_{kl} = \begin{cases} \mathbf{I}, & k = l \\ -\mathbf{J}^{k-l}, & k \neq l \end{cases}$$

Since $\mathbf{A}$ is rank deficient, $\mathbf{g}$ may be obtained from its singular value decomposition (SVD), as the singular vector corresponding to its smallest singular value. The transmitted symbols are then estimated from the effective inverse channel matrix $\mathbf{G}$ which is easily reconstructed from $\mathbf{g}$.

The variance σ_v^2, of the additive noise, which appears in the matrix $\mathbf{A}$ remains to be estimated. From (2), it is easy to see that

$$\mathbf{R}^{(x)}_{n-k,n-l} = \mathbf{H}\mathbf{R}^{(s)}_{n-k,n-l}\mathbf{H}^H + \sigma_v^2\mathbf{J}^{k-l} \tag{10}$$

so that $\mathbf{R}^{(x)}_{n-k,n-k} = \mathbf{H}\mathbf{R}^{(s)}_{n-k,n-k}\mathbf{H}^H + \sigma_v^2\mathbf{I}$. A sufficiently satisfactory estimate of σ_v^2 may be obtained as the smallest singular value in an SVD of the rank deficient matrix, $\mathbf{R}^{(x)}_{n-k,n-k}$ for any k. It must be noted that in contrast with [5, 10] we do not need an estimate of the dimension of the signal subspace. Obtaining a reliable estimate of the latter is, in general, nontrivial, and the robustness of many known techniques, is critically dependent on the same, implying thereby, that our approach enjoys a clear advantage in this context.

Thus far, our formulation has been inherently statistical, i.e., a practical implementation would assume that the received signal is time-ergodic and require the computation of temporally averaged autocorrelations for the ensemble averaged expectations that appear in the matrix $\mathbf{A}$. Since a large number of symbols is thus required to obtain reliable estimates of these autocorrelations, it follows that the resulting method would not be significantly faster than currently existing statistical methods [5, 10, 11].

Noting now that the relations (8)-(9) which define our method, do not appeal to the stationarity of the signal, we seek to use only short-term temporally averaged analogues of the ensemble averaged autocorrelations, i.e., we replace $\mathbf{R}^{(x)}_{n-k,n-l}$ in the matrix $\mathbf{A}$, by $\sum_{0\leq n\leq P-1} \mathbf{x}_{n-k}\mathbf{x}^H_{n-l}$, where P, the number of received symbols used, is at least LR (the product of the length of the modelled multipath channel times the number of antenna receivers at the base station) but is typically much less than that required to approximate the corresponding ensemble averaged expectation. The block hermitian matrix $\mathbf{A}$, then is no longer block toeplitz, but the validity of our development is retained. We thus expect to obtain good estimates of the effective inverse channel response with only a few received symbols, which are used to compute the short term time-averaged quantities that effectively replace the autocorrelations in the blocks that constitute $\mathbf{A}$, as explained above. This lends a deterministic angle to our approach, in that we make no assumptions regarding the statistics of the received signal, in the estimation process, besides using these short term time-averaged quantities that would correspond to ensemble averaged autocorrelations if a sufficiently large number of symbols were used.

It may nevertheless be argued that estimating the variance σ_v^2, of the additive noise makes the technique partially statistical. This is indeed true. Furthermore, it is obvious that the greater the number of received symbols used in computing the autocorrelations that constitute the matrix $\mathbf{R}^{(x)}_{n-k,n-k}$, an SVD of which is used to estimate σ_v^2, the better the estimate obtained. So there appears to be a tradeoff between the speed of the estimation process and its robustness in the presence of additive noise. We explain in what follows, how our approach strikes a balance between achieving both speed and robustness.

In the absence of additive noise, we have $\mathbf{A} = \mathbf{Q}$. Solving $\mathbf{Qg} = \mathbf{0}$ for $\mathbf{g}$ leads to perfect sequence estimation in an additive noise free environment, i.e., our scheme possesses the zero-forcing estimation property. When however, additive noise is present, as is usually the case, we note that its statistics do not vary anywhere as rapidly as the characteristics of even a slowly fading multipath wireless channel. Hence in practice, it is sufficient to compute estimates of σ_v^2, far less often than it is necessary to estimate $\mathbf{g}$. Our technique thus acquires a quasi-deterministic nature: it is of necessity, statistical insofar as the assumptions made in estimating the variance of the additive noise, but is deterministic and independent of the statistics of the received sequence otherwise. The fact that the additive noise variance needs to be estimated rather infrequently lends our technique the sought-after advantages of both speed and robustness.

The above scheme may be simplified (replacing **A** by a block tridiagonal matrix), by noting from (6) that

$$\begin{aligned}\mathbf{g}_{r-1}(\mathbf{R}^{(x)}_{n-M+r,n-M+r+1}+\sigma_v^2\mathbf{J}) &= \mathbf{g}_r(\mathbf{R}^{(x)}_{n-M+r+1,n-M+r+1}+\sigma_v^2\mathbf{I}) \\ &= \mathbf{g}_{r+1}(\mathbf{R}^{(x)}_{n-M+r+1,n-M+r+2}+\sigma_v^2\mathbf{J})^H \\ & \quad 0\le r\le M-1\end{aligned} \tag{11}$$

This may then be recast as

$$\begin{aligned}&\mathbf{g}_r(\mathbf{R}^{(x)}_{n-M+r+1,n-M+r+1}+\sigma_v^2\mathbf{I}) \\ &= \begin{cases}\mathbf{g}_{r+1}(\mathbf{R}^{(x)}_{n-M+r+1,n-M+r+2}+\sigma_v^2\mathbf{J})^H, & r=0\\ \frac{1}{2}(\mathbf{g}_{r-1}(\mathbf{R}^{(x)}_{n-M+r,n-M+r+1}+\sigma_v^2\mathbf{J})+\mathbf{g}_{r+1}(\mathbf{R}^{(x)}_{n-M+r+1,n-M+r+2}+\sigma_v^2\mathbf{J})^H), & 1\le r\le M-2\\ \mathbf{g}_{r-1}(\mathbf{R}^{(x)}_{n-M+r,n-M+r+1}+\sigma_v^2\mathbf{J}), & r=M-1\end{cases}\end{aligned} \tag{12}$$

In matrix form, the above reads as (8) where **A** is now a block tridiagonal matrix, with the block in position (k,l), is given by

$$(\mathbf{A})_{kl}=\begin{cases}2(\mathbf{R}^{(x)}_{n-M+k,n-M+k}+\sigma_v^2\mathbf{I}), & 1\le k=l\le M-2\\ (\mathbf{R}^{(x)}_{n-M+k,n-M+k}+\sigma_v^2\mathbf{I}), & k=l=0 \text{ or } M-1\\ -(\mathbf{R}^{(x)}_{n-M+l,n-M+k}+\sigma_v^2\mathbf{J}^{k-l}), & |k-l|=1\\ 0, & \text{otherwise}\end{cases} \tag{13}$$

The matrix on which the SVD must be performed, is now block tridiagonal, so that a more economical realisation becomes possible. Simulation studies show that this simplification does not lead to any degradation in the performance of our technique. All the results presented in Section 6 use this block tridiagonal form for the matrix **A**, in (8).

4 Symbol Identifiability Conditions

The necessary conditions for the existence of the effective inverse channel matrix **G** (presented in Section 2) are easily met in our case by choosing a sufficiently large data block length, and by the use of antenna array receivers [9]. A sufficient condition for the identifiability of **G** to the best possible extent stipulates that the signal subspace matrix have full row rank. This is summarised in the following theorem.

Theorem 2: The effective inverse channel response matrix **G** of a single-input R-output system with a block sylvester channel response matrix **H**, is uniquely identifiable upto a complex scaling factor i.e., $\mathbf{G}^H\mathbf{H}=\lambda\mathbf{I}_M$, if the signal subspace matrix $\mathbf{S}_n$ has full row rank.

Proof: See Appendix A

The extent to which **G** remains undetermined by the complex quantity λ, is essentially that of an arbitrary phase ambiguity, inherent in all such problems that require estimating a mixture matrix.

Our sufficiency requirement is obvious if the problem is viewed as seeking conditions for the factorisability of the received signal matrix $\mathbf{X}_n$ into a block sylvester matrix that have full column rank (i.e., the channel matrix **H**) and the transmitted signal matrix, i.e., $\mathbf{S}_n$. In this context the latter must clearly have full row rank, as noted in [15]. This condition on signal subspace matrix $\mathbf{S}_n$, which is block hankel, is quite mild, and is typically satisfied in digital communication systems of interest.

5 Computational Requirements

The computational requirements of several popular direct sequence estimation (or indirect channel estimation) techniques are compared with that of ours, in the table below.

Technique	Tong *et al*[5]	Moulines *et al* [10]	Liu and Xu [14]	Present work
# SVDs	3	2	2	2
Matrix sizes for SVDs	$NR \times NR$, $M \times M$, $M \times M$	$NR \times NR$, $LR \times LR$, (also one $M \times M$ matrix inversion)	$NR \times (m - M - 1)$, $M(m - 2M - 3) \times (M + NR)$	$NRM \times NRM$, $NR \times NR$

Table 1: A comparison of the computational requirements of blind estimation techniques (Note that $M = N + L - 1$ and m is the number of symbols being estimated)

From the above it is clear that our sequence estimation technique has the lowest computational requirements. In particular, the complexity of the algorithm in [14], depends on the number of symbols being detected, rather than on just a fixed block length of received data, making it a poor candidate for adaptive operation in time-varying propagation environments.

6 Performance Studies

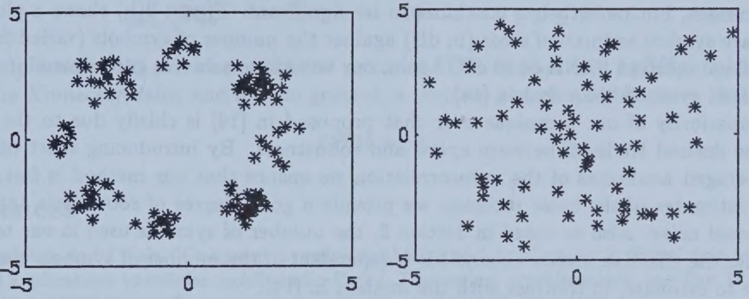
The performance of our technique was compared with that reported by Liu and Xu [14], for an 8-PSK source transmitting i.i.d. symbols over a three ray multipath reverse link wireless channel truncated to four symbol periods. The delays of the second and third rays with respect to the first were half and 1.2 times the symbol period. The signal (essentially a raised cosine waveform with 11 % rolloff as in [14]) from a single transmitting source, was spatially sampled by a four-antenna array base station receiver, so that there were four subchannels each of duration four symbol periods (i.e., R=4, L=4), with impulse response responses

$$\begin{aligned}
h_n^{(0)} &= (0.055 + 0.017i \quad -1.345 - 0.452i \quad 1.007 + 1.152i \quad 0.348 + 0.315i) \\
h_n^{(1)} &= (0.113 + 0.193i \quad -0.406 + 0.337i \quad 0.050 - 0.199i \quad 0.211 + 0.291i) \\
h_n^{(2)} &= (-0.808 - 0.318i \quad -0.431 - 0.261i \quad 1.282 + 1.146i \quad -0.361 - 0.274i) \\
h_n^{(3)} &= (-0.059 + 0.570i \quad -0.387 - 0.149i \quad 0.362 + 0.156i \quad -0.082 + 0.036i)
\end{aligned}$$

The received sequence was used in blocks of length equal to the maximum subchannel duration, i.e., $L = 4$.

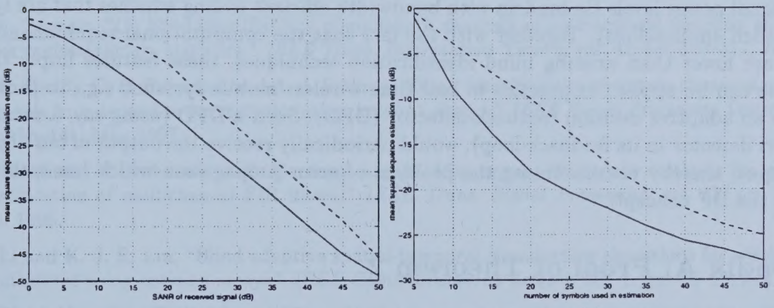
We examined the estimated constellations obtained from these two approaches, the variation of the mean square symbol estimation error with the received signal to additive noise ratio (SANR) for a fixed number of symbols used in estimation, and the variation of the mean square symbol estimation error with the number of symbols used in estimation for a given SANR. To allow a specific comparison of the gains accrued from our approach towards the estimation of the effective inverse channel, against the method in [14], we did not use any statistical criterion (e.g., minimum mean square error) in estimating $\mathbf{S}_n$, in our simulations, i.e., we ignored the effect of $\mathbf{V}_n$ in the final step of obtaining $\mathbf{S}_n$, given our estimate of $\mathbf{G}$.

Figure 1(a) shows the estimated constellation with our technique using 20 symbols, when the received SANR is 20 dB, while Figure 1(b) shows the same with the technique in [14]. Clearly both the number of symbols used and the received SANR are too low for the method in [14] (which relies solely on the algebraic properties of the signal subspace) to be able to acceptably detect the transmitted signal constellation in each case. Our technique however, demonstrates good performance, in that it successfully estimates the received sequence under the above conditions.



(a) (b)

Figure 1: Estimated symbol constellations with (a) Our technique, (b) Liu and Xu's technique [14], using 20 symbols at a received SANR of 20 dB



(a) (b)

Figure 2: The variation of the mean square symbol estimation error with (a) received SANR using only 20 symbols, and (b) the number of symbols used for estimation at a received SANR of 20 dB (Solid curve: Our technique; Dashed curve: Liu and Xu's technique [14])

Figure 2(a) shows the variation of the mean square symbol estimation error (in dB) when the received signal to additive noise ratio (SANR) is varied from 5 dB to 50 dB, using only 20 symbols for sequence estimation. The results shown were averaged over 100 independent Monte Carlo trials. Clearly our technique is significantly better than that in [14] for lower SANRs. The difference in the the estimation error between these two methods decreases as the received SANR increases, but nevertheless continues to be significant. Figure 2(b) shows a plot of the mean square symbol estimation error (in dB) against the number of symbols (varied from 5 to 50), for a fixed received SANR of 20 dB. Again, our technique achieves good detection with far fewer received symbols than that in [14].

The superiority of our technique over that proposed in [14] is chiefly due to the fact we achieve the desired trade-off between speed and robustness. By introducing short term temporally averaged analogues of the autocorrelation we ensure that our method is fast, and by including estimates of the noise variance, we provide a good degree of robustness against additive channel noise. Also as noted in Section 5, the number of symbols used in our technique (to estimate the effective inverse channel) is independent of the number of symbols that we are hence able to estimate, in contrast with the method in [14].

7 Conclusions

We have proposed a fast, robust, quasi-deterministic second order blind technique for sequence estimation in a reverse link wireless channel. This technique, described above for a single transmitting source, is easily generalised to a multiuser narrowband wireless mobile environment, where cochannel and adjacent channel interference are also suppressed. Our approach signifies an important contribution to achieving an optimal tradeoff between the speed of the estimation process and its robustness to increasing additive channel noise. In practical systems, where the signal power at the transmitting mobile is, of necessity, lower than that considered in this paper, our scheme may be initialised by the use of pilot symbols, implying semi-blind operation. Also, as noted earlier, a minimum mean square error, maximum likelihood, or maximum *a posteriori* probability criterion would typically be used in practice, to estimate the transmitted sequence from the effective inverse channel computed by our technique, further improving performance at low signal power levels (in tandem with bandwidth efficient coding schemes that are typically used in such applications). Together with the fact that the computational requirements of our scheme are lower than existing blind identification techniques, these features imply that our technique can be applied in practice to real time wireless mobile systems, e.g., to provide the input to an adaptive decision feedback detector (DFD). Such a DFD (using say, a maximum *a posteriori* detector in its feedback loop), would periodically receive the output of our algorithm, as its input, thereby circumventing the problem of error propagation which has traditionally plagued the DF concept.

Appendix A: Proof of Theorem 2

If the $M \times M$ hankel signal subspace matrix $\mathbf{S}_n$ (recall that $M = N + L - 1$) has full row rank, it has M linearly independent rows, so that $\mathbf{S}_n$ and $\mathbf{S}_{n-j}$ together have a total of $(M+j)$ linearly independent rows. Now, let $\mathbf{g}_m, 0 \leq m \leq M-1$, be a sequence of column vectors of length NR, such that

$$\mathbf{g}_n^H(\mathbf{x}_n - \mathbf{v}_n) \;=\; \mathbf{g}_{n-j}^H(\mathbf{x}_{n-j} - \mathbf{v}_{n-j}) \;\neq\; 0 \quad 1 \leq j \leq m \tag{14}$$

Then using the hankel forms of the received sequence and additive noise matrices, $\mathbf{X}_n$ and $\mathbf{V}_n$ respectively, we get $\mathbf{g}_m^H(\mathbf{X}_n - \mathbf{V}_n) = \mathbf{g}_{m-j}(\mathbf{X}_{n-j} - \mathbf{V}_{n-j})$, and hence

$$(\mathbf{g}_m^H\mathbf{H})\mathbf{S}_n \;=\; (\mathbf{g}_{m-j}^H\mathbf{H})\mathbf{S}_{n-j} \tag{15}$$

Using the fact that there are only $(M+j)$ linearly independent rows between $\mathbf{S}_n^H$ and $\mathbf{S}_{n-j}^H$, we see that (15) implies

$$\begin{array}{lcl} (\mathbf{g}_m^H\mathbf{H})_p & = & 0, \quad 0 \le p \le j-1, \\ (\mathbf{g}_m^H\mathbf{H})_{j+p} - (\mathbf{g}_{m-j}^H\mathbf{H})_p & = & 0, \quad 0 \le p \le M-j-1, \\ (\mathbf{g}_{m-j}^H\mathbf{H})_p & = & 0, \quad M-j \le p \le M-1, \end{array} \tag{16}$$

where $(\mathbf{v})_p$ denotes the pth element of the vector $\mathbf{v}$. This implies that $(\mathbf{g}_m^H\mathbf{H})_p = \lambda\delta_{mp}$, where δ_{mp} is the *Kronecker delta*, and λ is, in general, a nonzero complex constant, so that

$$\mathbf{G}^H\mathbf{H} = \lambda\mathbf{I}_M \tag{17}$$

References

[1] P. Balaban and J. Salz, "Optimum diversity combining and equalization in digital data transmission with applications to cellular mobile radio-Part I: Theoretical considerations, and Part II: Numerical results," *IEEE Trans. Communications*, vol. 40, pp. 885–907, May 1992.

[2] S. Bellini, "Blind equalization," *Alta Frequenza*, vol. 57, pp. 445–450, 1988.

[3] B. Porat and B. Friedlander, "Blind equalization of digital communication channels using higher-order moments," *IEEE Trans. Signal Processing*, vol. 38, 1991.

[4] W. A. Gardner, "A new method of channel identification," *IEEE Trans. Communications*, vol. 39, June 1991.

[5] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second order statistics: A time domain approach," *IEEE Trans. Information Theory*, vol. 40, pp. 340–349, Mar. 1994.

[6] L. Tong, G. Xu, B. Hassibi, and T. Kailath, "Blind identification and equalization based on second order statistics: A frequency domain approach," *IEEE Trans. Information Theory*, vol. 41, pp. 329–334, Jan. 1995.

[7] Y. Li and Z. Ding, "Blind channel identification based on second order cyclostationary statistics," in *Proc. Int. Conf. Accoustics, Speech, and Signal Processing*, pp. IV:81–84, Apr. 1993.

[8] J. K. Tugnait, "On blind identifiability of multipath channels using fractional sampling and second order cyclostationary statistics," *IEEE Trans. Information Theory*, vol. 41, pp. 308–311, Jan. 1995.

[9] V. U. Reddy, C. B. Papadias, and A. J. Paulraj, "Blind identifiability of certain classes of multipath channels from second-order statistics using antenna arrays," *IEEE Signal Processing Letters*, vol. 4, pp. 138–141, May 1997.

[10] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Trans. Signal Processing*, vol. 43, pp. 516–525, Feb. 1995.

[11] Y. Li and K. J. R. Liu, "Blind adaptive spatial-temporal equalization algorithms for wireless communications using antenna arrays," *IEEE Communications Letters*, vol. 1, pp. 25–27, Jan. 1997.

[12] M. Kristensson and B. Ottersten, "A statistical approach to subspace based blind identification," *IEEE Trans. Signal Processing*, vol. 46, pp. 1612–1623, June 1998.

[13] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least squares approach to blind channel identification," *IEEE Trans. Signal Processing*, vol. 43, pp. 2982–2993, Dec. 1995.

[14] H. Liu and G. Xu, "Closed-form blind symbol estimation in digital communications," *IEEE Trans. Signal Processing*, vol. 43, pp. 2714–2723, Nov. 1995.

[15] A. van der Veen, S. Talwar, and A. Paulraj, "A subspace approach to blind space-time signal processing for wireless communication systems," *IEEE Trans. Signal Processing*, vol. 45, pp. 173–190, Jan. 1997.

An Information-Theoretic Derivation of the MMSE Decision-Feedback Equalizer

Tommy Guess and Mahesh K. Varanasi

Department of Electrical and Computer Engineering
University of Colorado, Boulder, CO 80309-0425

****

Thirty-Sixth Annual Allerton Conference on Communication, Control, and Computing
September 23-25, 1998

Abstract

We consider the discrete-time Gaussian channel with inter-symbol interference (ISI). Under the assumption of perfect feedback, an information-theoretic derivation of the minimum mean-squared error (MMSE) decision-feedback equalizer (DFE) is presented. Whereas previous works have used information theory to analyze the zero-forcing and MMSE DFE structures, this work *derives* the MMSE DFE directly by means of information-lossless projections applied to the average mutual information of the Gaussian ISI channel. Specifically, the perfect-feedback MMSE DFE works by performing an information-lossless conversion of the ISI channel, which must be viewed in a sequence-wise manner, into a sequence of input-output pairs that must be viewed symbol-wise. With ideal interleaving, the latter can also be viewed as a memoryless Gaussian channel. This equivalence resolves the paradoxical result that perfect post-cursor ISI cancellation is in general an information-increasing operation. We find that it is not perfect cancellation that increases the mutual information, but rather the mathematically inconsistent assumption that the perfect-feedback MMSE DFE can be viewed as a sequence-wise channel with memory.

I. INTRODUCTION

Consider the discrete-time Gaussian inter-symbol interference (ISI) channel where the output of the channel at the k^{th} time instance is given as

$$y_k = \sum_{j=-\infty}^{\infty} a_j x_{k-j} + n_k. \tag{1}$$

The input sequence $\{x_k\}$ and the noise sequence $\{n_k\}$ are both complex-valued, zero-mean wide-sense stationary (WSS) processes that are statistically independent of each other. Moreover, the noise sequence is assumed to be circularly symmetric[1] and Gaussian. The sequence $\{a_k\}$ is assumed to be in ℓ_2, i. e., $\sum_{k=-\infty}^{\infty} |a_k|^2 < \infty$, so the the output sequence $\{y_k\}$ is also

This work was supported in part by NSF grants NCR-9406069 and NCR-9706591.

---

[1] A random variable is circularly symmetric or proper if its *pseudo* variance, i. e., $E[(x - E[x])^2]$, as opposed to $E[|x - E[x]|^2]$, is equal to zero. Similarly a discrete-time random process is circularly symmetric or proper if $E[x_{k+l}x_k] = 0$ for all k and l. See [1, Sec. 8.1.1] and [2]. We will assume throughout this paper that all Gaussian variables and processes are complex-valued and circularly symmetric.

WSS. This channel arises naturally in the context of quadrature amplitude modulation (QAM) signaling over a time-dispersive Gaussian channel when the received signal is processed by a matched filter and sampled at the symbol rate (see, for example, [1, Sec. 6.2.1] [3, Sec. 6.2] [4]).

To facilitate the presentation, we will formally use the z-transform to represent a convolution operator. Thus for any sequence $\{h_k\}$, stochastic or deterministic, we write that

$$h(z) = \sum_{k=-\infty}^{\infty} h_k z^{-k}. \tag{2}$$

This allows the ISI channel in (1) to be expressed as

$$y(z) = a(z)x(z) + n(z). \tag{3}$$

Of course, $x(z)$, $n(z)$, and $y(z)$ do not exist as true z-transforms, but since $\{x_k\}$, $\{n_k\}$, and $\{y_k\}$ are WSS processes, their auto- and cross-correlation functions do possess z-transforms. For example, if we let $R_{xy}(k) = E[x_{l+k}y_l^*]$, then the cross spectrum of $x(z)$ and $y(z)$ is given by

$$S_{xy}(z) = \sum_{k=-\infty}^{\infty} R_{xy}(k) z^{-k}. \tag{4}$$

In fact, it is easy to see that

$$\begin{aligned} S_y(z) &= a(z)S_x(z)a^*(1/z^*) + S_n(z) & (5)\\ S_{yx}(z) &= a(z)S_x(z), & (6)\end{aligned}$$

since $x(z)$ and $n(z)$ are independent. We have used the notation $a^*(1/z^*)$ to represent $\sum_{k=-\infty}^{\infty} a_k^* z^k$.

The average mutual information of the ISI channel, which we denote by $I(x(z); y(z))$, is given by [5] [6] [7]

$$I(x(z); y(z)) = \lim_{N\to\infty} \frac{1}{N} I(x_1^N; y_1^N), \tag{7}$$

where we have used the notation

$$I(x_1^N; y_1^N) = I(x_1, x_2, \ldots, x_N; y_1, y_2, \ldots, y_N) \tag{8}$$

with the function $I(\cdot\,;\cdot)$ denoting mutual information.

If we are given that $E[|x_k|^2] = p$, then it is well known that the particular input process $\{x_k\}$ that maximizes the average mutual information of the Gaussian ISI channel will itself be a Gaussian process [5]. The spectrum $S_x(z)$ that maximizes the average mutual information is a function of both the channel $a(z)$ and the noise spectrum $S_n(z)$, and it is determined by means of the so-called "water filling" procedure. Thus, we will assume hereafter that $\{x_k\}$ is a Gaussian process. This being the case, it is known that

$$I(x(z); y(z)) = \int_{-\pi}^{\pi} \log\left(\frac{S_y(e^{j\theta})}{S_n(e^{j\theta})}\right) \frac{d\theta}{2\pi} \tag{9}$$

in nats per two dimensions, since the symbols are complex valued.

In the sections that follow, Section II reviews some important results from both the information theory of Gaussian processes and prediction theory. These properties are incorporated in Section III to derive the MMSE DFE directly from $I(x(z); y(z))$. This approach is in contrast to previous works that have started with various DFE structures and then analyzed them in an information-theoretic context [8] [1, Sec. 10.5.5] [9] [10] [11]. There are distinct advantages to the approach contained herein. First, it gives insight as to how the MMSE DFE functions. The MMSE DFE receiver essentially converts the ISI channel into a sequence of input-output pairs that must be viewed symbol wise; with ideal interleaving, this can be viewed as a memoryless Gaussian channel. This conversion is achieved by means of information-lossless projections that yield a set of sufficient statistics. Additionally, we are able to resolve the paradoxical idea that perfect feedback in a MMSE DFE receiver structure can actually lead to a "fictitious" increase in mutual information. This idea is reported and discussed in [11], [8], and [10]. For perfect post-cursor ISI cancellation to be information increasing, the effective channel yielded by the MMSE DFE receiver must be viewed as a sequence-wise channel with memory, something our derivation shows is mathematically inconsistent. This resolution allows us to state a new converse to the coding theorem for the perfect-feedback MMSE DFE receiver.

II. SOME KEY RESULTS FROM INFORMATION THEORY AND LINEAR PREDICTION

In this section we present several results from information theory and prediction theory that are used to derive the MMSE DFE in Section III.

A. The Mutual Information Between Gaussian Processes

Here we are concerned with the mutual information between Gaussian random variables and processes. Some of the results we review are well known, but others are not.

Lemma 1: Let $X = \{x_k\}_{k\in\mathcal{G}}$, $Y = \{y_k\}_{k\in\mathcal{H}}$, $Z = \{z_k\}_{k\in\mathcal{J}}$, and (X, Y, Z) each be a set of jointly Gaussian zero-mean random variables with the sets $\mathcal{G}$, $\mathcal{H}$, and $\mathcal{J}$ being arbitrary. Then the conditional mutual information

$$I(X; Y|Z) = I((\mathrm{E} - \Pi_Z)(X); (\mathrm{E} - \Pi_Z)(Y)), \tag{10}$$

where E denotes the identity operator and Π_Z denotes the operator that orthogonally projects on to Z. ◇

This lemma is shown in [12, Sec. 9.3]. Since the operator $(\mathrm{E} - \Pi_Z)$ yields that part of the argument which is orthogonal to Z, we see that the conditional mutual information has been expressed as the mutual information between the parts of X and Y that are orthogonal to Z.

Lemma 2: With the same hypotheses as Lemma 1, let Π_Y be the orthogonal-projection operator onto Y. Then we have that

$$I(X;Y) = I(X;\Pi_Y(X)). \tag{11}$$

◇

The previous two lemmas follow naturally from the more fundamental result for Gaussian variables that the mutual information between X and Y is a function of only the angles between their associated Hilbert spaces [12, Sec. 9.2] [13]. For example, in Lemma 2 the non-zero angles between X and Y are the same as those between X and $\Pi_Y(X)$ so that the mutual information is unaltered. And in Lemma 1, those parts of X and Y that lie in Z contribute nothing to the mutual information. The inner product of the Hilbert space is expectation, so for zero-mean Gaussian random variables x and y we have that

$$\Pi_y(x) = E[x|y] = \frac{E[xy^*]}{E[|y|^2]}y. \tag{12}$$

The following lemma is a variation of the well known log(1 + SNR) result for the mutual information between two zero-mean Gaussian random variables, i. e., the capacity of memoryless discrete-time Gaussian channel; it evaluates the signal-plus-noise power divided by the noise power:

Lemma 3: Given zero-mean jointly Gaussian random variables x and y,

$$I(x;y) = \log\left(\frac{E[|y|^2]}{E[|y - E[y|x]|^2]}\right). \tag{13}$$

◇

Lemma 4: Suppose that at least one of the zero-mean WSS processes $\{x_k\}$ and $\{y_k\}$ is regular (i. e., non-predictable). That is, $\int_{-\pi}^{\pi}|\log(S_x(e^{j\theta}))|d\theta < \infty$ or $\int_{-\pi}^{\pi}|\log(S_y(e^{j\theta}))|d\theta < \infty$. Then for all k we have that

$$I(x(z);y(z)) = I(x_k;y_{-\infty}^{\infty}|x_{-\infty}^{k-1}). \tag{14}$$

◇

For a discussion of this result see [12, Thm. 10.2.1].

B. *Results from Linear Prediction Theory*

Suppose we are given the regular (cf. Lemma 4), zero-mean WSS process $\{x_k\}$. The innovations process for $\{x_k\}$ is given by

$$(i_x)_k = x_k - E[x_k|x_{-\infty}^{k-1}]. \tag{15}$$

From linear prediction theory for scalar processes, we know that

$$x_k = \sum_{j=0}^{\infty}(\phi_x)_j(i_x)_{k-j}, \tag{16}$$

where $\phi_x(z) = \sum_{k=0}^{\infty}(\phi_x)_k z^{-k}$ is causal, monic (i. e., $(\phi_x)_0 = 1$), and minimum-phase [14, Secs.10-5, 13-4]. This means that $(\phi_x(z))^{-1}$, which we shall write as $\phi_x^{-1}(z)$, is also causal and monic. Moreover, we have that

$$S_x(z) = g_x \phi_x(z)\phi_x^*(1/z^*), \tag{17}$$

where $g_x = E[|(i_x)_k|^2]$ is the geometric mean of the spectrum $S_x(e^{j\theta})$. That is,

$$g_x = \exp\left(\int_{-\pi}^{\pi} \log(S_x(e^{j\theta}))\frac{d\theta}{2\pi}\right). \tag{18}$$

This allows us to formally write that $x(z) = \phi_x(z)i_x(z)$, and if we consider the formal z-transform of the sequence $\{\hat{x}_k\} = \{E[x_k|x_{-\infty}^{k-1}]\}$, we get that $\hat{x}(z) = x(z) - i_x(z) = (1 - \phi_x^{-1}(z))x(z)$. These results are summarized in the following lemma.

Lemma 5: Given a regular, zero-mean WSS process $\{x_k\}$, define the processes $\{\hat{x}_k\}$ and $\{(i_x)_k\}$ by $\hat{x}_k = E[x_k|x_{-\infty}^{k-1}]$ and $(i_x)_k = x_k - \hat{x}_k$. Then

$$\begin{aligned} x(z) &= \phi_x(z)i_x(z) && (19)\\ i_x(z) &= \phi_x^{-1}(z)x(z) && (20)\\ \hat{x}(z) &= (1-\phi_x^{-1}(z))x(z). && (21) \end{aligned}$$

The polynomial $\phi_x(z)$ is causal, monic, and minimum phase; the spectral factorization of $S_x(z)/g_x$ is $\phi^*(1/z^*)\phi(z)$, where g_x given by (18). ◇

Having established the necessary background we proceed in the next section to derive the MMSE DFE receiver structure.

III. AN INFORMATION-THEORETIC DERIVATION OF THE MMSE DFE

We begin by proving the average mutual-information result previously stated in equation (9). Our approach parallels that given in [12, Sec. 10.2]. For the Gaussian ISI channel in (3) with a regular, zero-mean WSS Gaussian input, we consider $I(x(z); y(z))$. We know from Lemma 4 that

$$I(x(z); y(z)) = I(x_k; y_{-\infty}^{\infty}|x_{-\infty}^{k-1}) \tag{22}$$

for all k. Using the chain rule of mutual information, we can express $I(x_k; y_{-\infty}^{\infty}, x_{-\infty}^{k-1})$ in two different ways.

$$\begin{aligned} I(x_k; y_{-\infty}^{\infty}, x_{-\infty}^{k-1}) &= I(x_k; y_{-\infty}^{\infty}) + I(x_k; x_{-\infty}^{k-1}|y_{-\infty}^{\infty}) && (23)\\ I(x_k; y_{-\infty}^{\infty}, x_{-\infty}^{k-1}) &= I(x_k; x_{-\infty}^{k-1}) + I(x_k; y_{-\infty}^{\infty}|x_{-\infty}^{k-1}). && (24) \end{aligned}$$

Combining these two results with (22) allows us to see that

$$I(x(z); y(z)) = I(x_k; y_{-\infty}^{\infty}) + I(x_k; x_{-\infty}^{k-1}|y_{-\infty}^{\infty}) - I(x_k; x_{-\infty}^{k-1}). \tag{25}$$

Applying Lemma 2 to the first term and Lemma 1 to the second term we get

$$I(x(z); y(z)) = I(x_k; E[x_k | y_{-\infty}^{\infty}]) + I(x_k - E[x_k | y_{-\infty}^{\infty}]; \{x_j - E[x_j | y_{-\infty}^{\infty}]\}_{j=-\infty}^{k-1}) - I(x_k; x_{-\infty}^{k-1}). \tag{26}$$

Define the sequences $\{\tilde{x}_k\}$ and $\{e_k\}$ by $\tilde{x}_k = E[x_k | y_{-\infty}^{\infty}]$ and $e_k = x_k - \tilde{x}_k$. Making use of the hat-notation for the one-step predictor as in Lemma 5, i. e., $\hat{x}_k = E[x_k | x_{-\infty}^{k-1}]$ and $\hat{e}_k = E[e_k | e_{-\infty}^{k-1}]$, we find that

$$\begin{aligned} I(x(z); y(z)) &= I(x_k; \tilde{x}_k) + I(e_k; e_{-\infty}^{k-1}) - I(x_k; x_{-\infty}^{k-1}) && (27) \\ &= I(x_k; \tilde{x}_k) + I(e_k; \hat{e}_k) - I(x_k; \hat{x}_k). && (28) \end{aligned}$$

Thus, the average mutual information of the Gaussian ISI channel has been expressed in terms of three pairs of scalar-valued Gaussian variables. Before considering these pairs in more detail, we apply Lemmas 3 and 5 to (28) to get that

$$\begin{aligned} I(x(z); y(z)) &= \log\left(\frac{E[|x_k|^2]}{E[|x_k - \tilde{x}_k|^2]}\right) - \log\left(\frac{E[|x_k|^2]}{E[|x_k - \hat{x}_k|^2]}\right) + \log\left(\frac{E[|e_k|^2]}{E[|e_k - \hat{e}_k|^2]}\right) \\ &= \log\left(\frac{E[|x_k|^2]}{E[|e_k|^2]} \frac{E[|(i_x)_k|^2]}{E[|x_k|^2]} \frac{E[|e_k|^2]}{E[|(i_e)_k|^2]}\right) \\ &= \log\left(\frac{g_x}{g_e}\right). && (29) \end{aligned}$$

We can explicitly calculate $S_e(z)$ to be

$$\begin{aligned} S_e(z) &= S_x(z) - S_{xy}(z) S_y^{-1}(z) S_{yx}(z) && (30) \\ &= \frac{S_x(z) S_n(z)}{S_y(z)}, && (31) \end{aligned}$$

so that $g_e = g_x g_n / g_y$. Finally then, the result given in (9) is derived since $I(x(z); y(z))$ equals $\log(g_y / g_n)$.

We now look more closely at the three mutual-information terms that compose the right-hand side of (28). The first term is $I(x_k; \tilde{x}_k)$. Since $\tilde{x}(z)$ is the orthogonal projection of $x(z)$ onto $y(z)$, it can be expressed as $\tilde{x}(z) = S_{xy}(z) S_y^{-1}(z) y(z)$. Incorporating (5) and (6), we have that

$$\tilde{x}(z) = c(z)(a(z)x(z) + n(z)), \tag{32}$$

where

$$c(z) = \frac{S_x(z) a^*(1/z^*)}{a(z) S_x(z) a^*(1/z^*) + S_n(z)}. \tag{33}$$

We must be very careful, however, in our interpretation of this representation. Without further clarification, the channel in (32) represents $I(x(z); \tilde{x}(z))$, not $I(x_k; \tilde{x}_k)$. These are two very distinct channels. The first is a sequence-wise interpretation of the channel, while the second is a symbol-wise interpretation of the channel. Clearly, only this second interpretation is relevant to our discussion.

The second term in (28) is $I(e_k; \hat{e}_k)$. By Lemma 5, this is representable as

$$\hat{e}(z) = (1 - \phi_e^{-1}(z))e(z), \tag{34}$$

where $\phi_e^{-1}(z)$ is the causal, monic, minimum-phase whitening filter for $S_e(z)/g_e$, and of course $e(z) = x(z) - \tilde{x}(z)$. Again, we are interested in only the k^{th} input and the k^{th} output. And similarly, the third channel $I(x_k; \hat{x}_k)$ comes from the following

$$\hat{x}(z) = (1 - \phi_x^{-1}(z))x(z). \tag{35}$$

It turns out that the inputs and outputs of the channels in (32), (34), and (35) are inter-related in such a manner that allows them to be combined to form a channel that is equivalent as far as mutual information is concerned. The first and second channels are combined as shown in parts (a) and (b) of Figure 1. Since we are adding the outputs of the two channels, we must still verify that the mutual information of the overall channel is equal to the sum of the mutual informations of the two component channels.

Lemma 6: When the channels $I(x_k; \tilde{x}_k)$ and $I(e_k; \hat{e}_k)$ are combined as shown in Figure 1, we have that

$$I(x_k; \hat{e}_k + \tilde{x}_k) = I(x_k; \tilde{x}_k) + I(e_k; \hat{e}_k). \tag{36}$$

This lemma can be proved directly. For example, see [10], [11], or [1, chap. 10] where the left-hand side is shown to equal $I(x(z); y(z))$ when $\{x_k\}$ is white. The result can also be shown to follow from a much more general statement concerning certain relationships of orthogonal-projection channels. The latter approach sidesteps the algebraic details associated with the direct proof. It also allows one to easily generalize the result to other Gaussian channels such as the multivariate ISI channel [15], the finite-dimensional ISI channel [16], and the synchronous Gaussian multiple-access channel [17].

It should be pointed out that channel shown in part (b) of Figure 1 is sometimes given as the perfect-feedback MMSE DFE structure for the Gaussian ISI channel [1, Sec. 10.2]. We see, however, that unless $\{x_k\}$ is a white process so that $I(x_k; \hat{x}_k) = 0$, this structure does not model the Gaussian ISI channel accurately. Equation (23) shows that it models the channel $I(x_k; y_{-\infty}^{\infty}, x_{-\infty}^{k-1})$ instead of the ISI channel of interest, $I(x_k; y_{-\infty}^{\infty} | x_{-\infty}^{k-1})$. To incorporate the third channel $I(x_k; \hat{x}_k)$, we can express $x(z)(1 - \phi_e^{-1}(z))$ as $x(z)(1 - \phi_x^{-1}(z)) + x(z)(\phi_x^{-1}(z) + \phi_e^{-1}(z))$. When this representation is incorporated into part (b) of Figure 1, the result is part (c) of the same figure. Note that the last parallel branch of this figure is is simply the third channel of interest. Thus, the subtraction of the third mutual-information term in (28) suggests the removal of this parallel branch. We know that

$$\phi_x^{-1}(z) - \phi_e^{-1}(z) = \phi_x^{-1}(z)\left(1 - \frac{\phi_y(z)}{\phi_n(z)}\right). \tag{37}$$

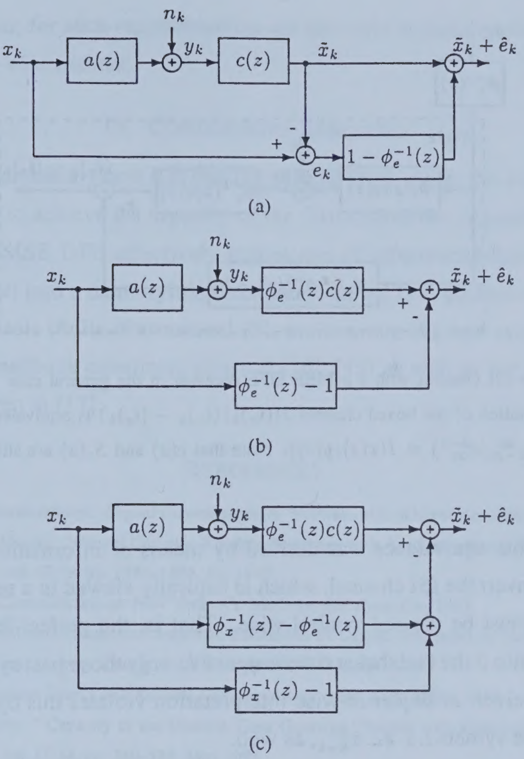


Fig. 1. Three equivalent forms of the channel $I(x_k; \tilde{x}_k + \hat{e}_k)$. Whenever $\{x_k\}$ is not white this is the same channel as $I(x_k; y_{-\infty}^{\infty}, x_{-\infty}^{k-1})$, but if $\{x_k\}$ is white then it is the same as the Gaussian ISI channel, i. e., $I(x_k; y_{-\infty}^{\infty} | x_{-\infty}^{k-1}) = I(x(z); y(z))$.

so that the combination of all three terms in (28) corresponds to the model pictured in Figure 2. It is easily seen that $\hat{e}_k + \tilde{x}_k - x_k + (i_x)_k = (i_x)_k - (i_e)_k$, and it is not hard to show that $I((i_x)_k; (i_x)_k - (i_e)_k) = I(x(z); y(z))$. In addition, we can show that $I(x_k; (i_x)_k - (i_e)_k) < I((i_x)_k; (i_x)_k - (i_e)_k)$ whenever $\{x_k\}$ is not white. Hence, Figure 2 only makes sense if we consider $(i_x)_k$ to be the input. Note that the symbol-wise interpretation of this channel is equivalent to the memoryless channel that is yielded by perfect interleaving. This, then, is the canonical MMSE DFE model, and it is the model that is typically assumed in the literature.[2]

Thus, we have shown that from the viewpoint of mutual information, there is no loss in generality in assuming that the receiver of the Gaussian ISI channel is the perfect-feedback MMSE DFE. Of course, perfect feedback means that the post-cursor ISI is completely removed

---

[2] The output $(i_x)_k - (i_e)_k$ is sometimes scaled by the constant α that yields the unbiased MMSE DFE, i. e., $E[\alpha((i_x)_k - (i_e)_k) | (i_x)_k] = (i_x)_k$ [10] [1, Sec. 10.2]. But as far as mutual information is concerned, there is no difference between the biased MMSE DFE and the unbiased MMSE DFE since $I((i_x)_k; \alpha[(i_x)_k - (i_e)_k]) = I((i_x)_k; (i_x)_k - (i_e)_k)$ for all non-zero α.

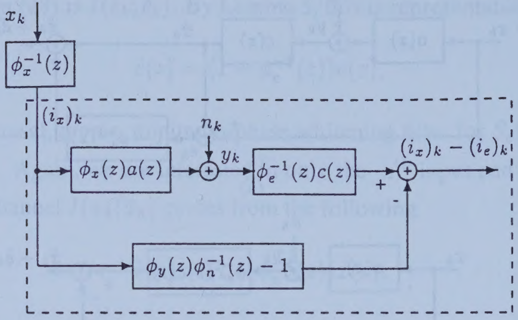


Fig. 2. The Gaussian ISI channel with a MMSE DFE receiver in the general case where $\{x_k\}$ is not white. The mutual information of the boxed channel $I((i_x)_k; (i_x)_k - (i_e)_k)$ is equivalent to that of the Gaussian ISI channel $I(x_k; y_{-\infty}^{\infty} | x_{\infty}^{k-1}) = I(x(z); y(z))$. Note that $c(z)$ and $S_e(z)$ are still given by (33) and (31), respectively.

by the receiver. This equivalence was derived by means of information-lossless orthogonal projections that convert the ISI channel, which is naturally viewed in a sequence-wise manner, to a channel that must be viewed symbol-wise. That is, the perfect-feedback MMSE DFE structure has built into it the fact that at time instance k, only those past symbols, i. e., $x_{-\infty}^{k-1}$, are available at the receiver. A sequence-wise interpretation violates this by inherently assuming knowledge of future symbols, i. e., x_{k+1}^{∞}, as well.

In [10, Prop. 1], the following direct part of the coding theorem for the perfect-feedback MMSE DFE is established.

Theorem 1 (Direct) Suppose we are given a Gaussian ISI channel with capacity C_{ISI}. If C_{DFE} denotes the capacity of the overall channel when the receiver is constrained to be the perfect-feedback MMSE DFE, then $C_{\text{DFE}} \geq C_{\text{ISI}}$. ◇

We are able to provide the converse to this theorem. That is,

Theorem 2 (Converse) $C_{\text{DFE}} \leq C_{\text{ISI}}$. ◇

This converse, was previously conjectured to be false because of the belief that perfect post-cursor ISI cancellation is in general an information-increasing operation [11] [8] [10]. This belief is based on the correctly observed fact that in general $I(i_x(z); i_x(z) - i_e(z)) > I((i_x)_k; (i_x)_k - (i_e)_k)$. But as we have indicated, these are two, very distinct channels. Our development has shown that the perfect-feedback MMSE DFE is a receiver structure that converts $I(x(z); y(z))$ into the latter. That is to say, the sequence of pairs $\{(i_x)_k; (i_x)_k - (i_e)_k\}$ when viewed symbol wise is sufficient for the pair $(x(z); y(z))$. The channel $I(i_x(z); i_x(z) - i_e(z))$ requires a sequence-wise interpretation of the MMSE DFE, an interpretation that is clearly incompatible with the MMSE DFE and, hence, also incompatible with the ISI channel. This highlights a potential danger of working with block diagrams of the Gaussian ISI channel with a MMSE DFE receiver, for such representations are not valid without the caveat that they *must* be viewed in symbol-wise manner.

IV. CONCLUDING REMARKS

We have seen fundamentally how it is that the MMSE DFE, under the assumption of perfect feedback, allows one to achieve the capacity of the Gaussian-noise channel with inter-symbol interference. The MMSE DFE effectively makes use of information-lossless projections to convert the ISI channel into a memoryless Gaussian channel. Our method of derivation can be applied to other channels (finite-dimensional ISI, multivariate ISI, and synchronous multiple-access) to yield the feedback equalizers given in [16], [15] as well as the multiuser decision-feedback receiver given in [17].

REFERENCES

[1] E. A. Lee and D. G. Messerschmitt, *Digital Communication.* Norwell, MA: Kluwer Academic Publishers, 1994.

[2] F. D. Neeser and J. L. Massey, " Proper Complex Random Processes with Applications to Information Theory ," *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 1293–1302, Jul. 1993.

[3] J. G. Proakis, *Digital Communications.* New York, NY: McGraw-Hill Book Co., 1983.

[4] G. D. Forney Jr., " Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference ," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 363–378, May 1972.

[5] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[6] W. Hirt and J. L. Massey, " Capacity of the Discrete-Time Gaussian Channel with Intersymbol Interference ," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 380–388, May 1988.

[7] S. Shamai (Shitz), L. H. Ozarow, and A. D. Wyner, " Information Rates for Discrete-Time Gaussian Channel with Intersymbol Interference and Stationary Inputs ," *IEEE Trans. Inform. Theory*, vol. IT-37, pp. 1527–1539, Nov. 1991.

[8] J. R. Barry, E. Lee, and D. G. Messerschmitt, " Capacity Penalty Due to Ideal Zero-Forcing Decision-Feedback Equalization ," *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 1062–1071, Jul. 1996.

[9] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and J. Forney, G. D., " MMSE Decision-Feedback Equalizers and Coding-Part I: Equalization Results ," *IEEE Trans. Commun.*, vol. COM-43, pp. 2582–2594, Oct. 1995.

[10] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney Jr., " MMSE Decision-Feedback Equalizers and Coding-Part II: Coding Results ," *IEEE Trans. Commun.*, vol. COM-43, pp. 2595–2604, Oct. 1995.

[11] S. Shamai (Shitz) and R. Laoria, " The Intersymbol Interference Channel: Lower Bounds on Capacity and Channel Precoding Loss ," *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 1388–1404, Sep. 1996.

[12] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes.* Holden-Day, Inc., 1964.

[13] I. M. Gelfand and A. M. Yaglom, " Calculation of the Amount of Information About a Random Function Contained in another Such Function ," *American Mathematical Society Translations*, vol. 12, pp. 199–246, 1959.

[14] A. Papoulis, *Probability, Random Variables, and Stochastic Processes, Second Edition.* New York, NY: McGraw-Hill Book Co., 1984.

[15] J. Yang and S. Roy, " Joint Transmitter-Receiver Optimization for Multi-Input Multi-Output Systems with Decision Feedback ," *IEEE Trans. Inform. Theory*, vol. IT-40, pp. 1334–1347, Sep. 1994.

[16] N. Al-Dhahir, J. M. Cioffi, and G. D. Forney Jr., " A Canonical Finite-Length MMSE-DFE," in *Proc. Thirty-third Allerton Conf. Commun., Contr. and Comput.*, Allerton, IL, pp. 598–607, Oct. 1995.

[17] M. K. Varanasi and T. Guess, " Optimum Decision Feedback Multiuser Equalization with Successive Decoding Achieves the Total Capacity of the Gaussian Multiple-Access Channel ," in *Proc. Thirty-First Asilomar Conf. Signals, Systems, and Computers* , Nov. 1997.

Optimized Decision Feedback Equalization for Convolutional Coding with Reduced Delay*

Jung-Tao Liu and Saul B. Gelfand
School of Electrical and Computer Engineering
Purdue University, W. Lafayette IN

Abstract

Error propagation is a significant problem with the DFE. In particular, when a DFE is concatenated with a convolutional code the burst errors associated with error propagation can severely degrade performance. Interleaving can be employed to break up the error bursts but it can introduce excessive delays. Here, we propose certain stationary error models and derive modified DFEs which can compensate for the error propagation so as to break up error bursts and improve performance with convolutional codes. The incorporation and removal of the bias during the design process is discussed. With coding, the modified DFEs can improve on the conventional DFE in terms of bit error rate, and the modified DFEs without interleaving can improve on the conventional DFE with interleaving in terms of decision delay. The modified DFE can be viewed as a type of joint equalization and interleaving technique.

1 Introduction

The decision feedback equalizer (DFE) is a relatively low complexity approach to combating intersymbol interference (ISI) in digital communication systems. The DFE suffers from error propagation, which can result in more than 2dB performance loss [1]. The DFE can be used for equalization in coded systems (operating at low to moderate SNR) provided the error propagation can be controlled and/or compensated. In these systems, the convolutional code is a good choice as it is optimized for the ISI-free channel. To combat the error propagation, the transmitted bits can be interleaved, and the DFE can be concatenated with a deinterleaver and decoder at the receiver. This separate equalization and decoding scheme is the simplest approach and requires the least (but possibly still substantial) interleaver depth. See [2] for a detailed discussion of combined equalization and decoding techniques, which offer improved performance but with increased delay and complexity.

In this paper we examine the idea of incorporating a stationary error model into the design of the DFE so as to can compensate for the effects of error propagation in convolutionally coded systems. This is to be contrasted with the analysis of error propagation with the (correct) Markov error propagation model [3], which seems intractable from the design viewpoint. We propose two stationary error models and give derivations of modified DFEs (MDFEs), which differ from the conventional DFE only in the values of the tap weights. This is different from other approaches which modify the structure (and complexity) of the DFE to mitigate error propagation (c.f. [4] and [5]). Our approach is most closely related to the recent work of [6], who also derived a modified DFE based on a (different) stationary error model, and gave some results which showed improvement in minimum mean-square error (MMSE) for uncoded systems. In the present work, we systematically investigate the structure of plausible stationary error models, and the

---

*This work was supported in part by the National Science Foundation under Grant NCR-9706372

derivation of modified DFEs which are based on them. It is known that the minimum mean-square error (MMSE) DFE is biased [7]. We show how to remove the bias during the design process, so that both the bias *and* error propagation are jointly compensated for. We also show how to incorporate the stationary error model into union bounds which hence crudely reflect the role of error propagation in DFE with convolutional coding. In our simulations it is found that there is little improvement in bit error rates (BER) with the MDFE relative to the DFE in the uncoded case, but dramatic improvement with some convolutional codes. This is shown to be due to the MDFE's ability to increase isolated errors in favor of burst errors, i.e., it has the effect of a joint equalizer and interleaver. We also compare the performance of the MDFE without interleaver to the DFE with (convolutional) interleavers of varying depth, and show that the DFE requires much longer overall delay to achieve the same performance as the MDFE.

2 Channel Model and DFE Structure

The equivalent discrete-time additive Gaussian noise channel shown in Figure 1a) is used throughout this work, where $\{I_k\}$ are i.i.d symbols from an $M-$PAM constellation taking values $\{\pm 1, \pm 3, ..., \pm(M-1)\}$ with equal probability, and variance σ_I^2; $\{x_k\}$ is the finite-length symbol-rate sampled response of the combined transmit filter, channel and (matched) receiver filter; and $\{\xi_k\}$ is Gaussian noise with mean 0 and autocorrelation function $r_{\xi\xi}(j) = E\{\xi_i \xi_{i-j}\} = \sigma_n^2 x_j$ (and $\{I_k\}$ and $\{\xi_k\}$ are independent). The channel and noise variance are assumed known and fixed in the following derivations.

The DFE is also shown in Figure 1a). The output of the DFE before slicing is denoted by $\hat{I}_k$, and after slicing by $\tilde{I}_k = dec(\hat{I}_k)$. For the finite impulse response (FIR) DFE we represent the feedforward and feedback filters by the vectors $\underline{\omega}^t = [\omega_{-N}, \dots, \omega_0]$, and $\underline{b}^t = [b_1, \dots, b_L]$, respectively, so that

$$\hat{I}_k = \underline{\omega}^t \underline{y}_k - \underline{b}^t \underline{\tilde{I}}_k = \underline{\omega}^t X \underline{I}_k + \underline{\omega}^t \underline{\xi}_k - \underline{b}^t \underline{\tilde{I}}_k$$

where $\underline{y}_k^t = \left[\begin{array}{ccc} y_{k+N}, & \cdots, & y_k \end{array}\right]$, $\underline{I}_k^t = \left[\begin{array}{ccc} I_{k+N+L}, & \cdots, & I_{k-L} \end{array}\right]$, $\underline{\tilde{I}}_k^t = \left[\begin{array}{ccc} \tilde{I}_{k-1}, & \cdots, & \tilde{I}_{k-L} \end{array}\right]$, $\underline{\xi}_k^t = \left[\begin{array}{ccc} \xi_{k+N}, & \cdots, & \xi_k \end{array}\right]$, and the channel matrix X is a $(N+1) \times (N+2L+1)$ matrix with elements: $[X]_{ij} = x_{-L+j-i}$, for $0 \le j-i \le 2L$, and zero otherwise. The number of feedback taps is set equal to the channel memory L; this "key assumption" (see [8]) is not essential but makes the analysis easier.

3 DFE with Error Propagation Compensation

Here we propose two stationary error models for the detected symbols, and then develop a procedure to iteratively compute modified DFE coefficients which compensate for the error. The stationary error models are constructed from the statistical properties of the conventional infinite impulse response zero-forcing DFE (IIR ZF-DFE) and the infinite impulse response minimum mean-square error DFE (IIR MMSE-DFE).

3.1 Zero-Forcing Error Propagation Model

Let the slicer input and output be written as $\hat{I}_k = I_k + e_k$ and $\tilde{I}_k = I_k + v_k$, respectively. The $\{e_k\}$ are the equalizer output errors, and the $\{v_k\}$ are the detected symbol errors. We start by examining the structure of the detected symbol error statistics for the IIR ZF-DFE. Assuming correct past decisions, the slicer input $\hat{I}_k$ can be written as $\hat{I}_k = I_k + e_k$ where $e_k = \sum_{n=-\infty}^{0} \omega_n \xi_{k-n}$. $\{e_k\}$ is a stationary independent zero-mean Gaussian sequence which is independent of $\{I_k\}$. Since the slicer is memoryless, v_k is a function of I_k and e_k only, and therefore the detected symbol errors $\{v_k\}$ are also a stationary independent zero-mean sequence, with v_k independent of $\{I_n, e_n\ n \neq k\}$. Now the IIR ZF-DFE can be viewed as an anticausal (and anticausally invertible) whitening filter, which whitens $\{\xi_k\}$ to $\{e_k\}$, as shown in Figure 2a) [7]. Hence e_k and so v_k is independent of $\{\xi_n, n > k\}$. We summarize the above in the following way:

1. $\{v_k\}$ is a stationary independent zero-mean, jointly stationary with $\{I_k\}$ and $\{\xi_k\}$.
2. v_k is independent of $\{I_n, n \neq k\}$ and $\{\xi_n, n > k\}$.

Note in particular that the second order statistics satisfy $r_{vv}(k) = 0\ \forall\ k \neq 0$, $r_{Iv}(k) = 0\ \forall\ k \neq 0$, and $r_{\xi v}(k) = 0\ \forall\ k > 0$.

We next incorporate this model into the design of the DFE, yielding a modified DFE based on a ZF error propagation model. The modified DFE is actually derived using the MMSE criterion, so as to incorporate the statistical properties of the error propagation model.

3.2 Modified DFE with ZF Error Propagation Model

For the MDFE, past decisions are not necessarily correct, and $\tilde{I}_k = I_k + v_k$ is input to the feedback filter. The output of such a DFE can thus be written as

$$\hat{I}_k = \underline{\omega}^t X \underline{I}_k + \underline{\omega}^t \underline{\xi}_k - \underline{b}^t \underline{I}_{k-} - \underline{b}^t \underline{v}_k, \tag{1}$$

where $\underline{v}_k^t = \left[\begin{array}{ccc} v_{k-1}, & \cdots & , v_{k-L} \end{array}\right]$ and $I_{k-} = \left[\begin{array}{ccc} I_{k-1}, & \cdots & , I_{k-L} \end{array}\right]$. For the ZF error propagation model in Section 3.1, the MSE can be expressed as,

$$\begin{aligned} MSE(\underline{\omega}, \underline{b}) &= \sigma_I^2 - \sigma_I^2 2 \underline{\omega}^t X \underline{u} - 2\left(\sigma_I^2 + r_{Iv}\right) \underline{\omega}^t X \left[\frac{0_{(N+L+1)\times L}}{I_{L\times L}}\right] \underline{b} \\ &\quad + \underline{\omega}^t \left[\sigma_I^2 X X^t + \sigma_n^2 \Phi\right] \underline{\omega} + \left(\sigma_I^2 + \sigma_v^2 + 2 r_{Iv}\right) \underline{b}^t \underline{b} \end{aligned} \tag{2}$$

where $\sigma_v^2 = E\{v_k^2\}$, $r_{Iv} = E\{I_k v_k\}$, $\underline{u}^t = \left[0_{1\times(N+L)}, 1, 0_{1\times L}\right]$, and Φ is a matrix with elements $\phi_{ij} = x_{j-i}$, $\forall\, |j-i| \leq L$, and $\phi_{ij} = 0$, otherwise. The optimal feedforward and feedback filters are found by minimizing this MSE, which gives

$$\begin{aligned} \underline{b}^t &= \frac{r_{I\tilde{I}}}{\sigma_{\tilde{I}}^2} \underline{\omega}^t X \left[\frac{0_{(N+L+1)\times L}}{I_{L\times L}}\right] \\ \underline{\omega}^t &= \sigma_I^2 \underline{u}^t X^t \left\{ X \left[\begin{array}{c|c} \sigma_I^2 \cdot I_{(N+L+1)} & 0_{(N+L+1)\times L} \\ \hline 0_{L\times(N+L+1)} & r_{I|\tilde{I}} I_L \end{array}\right] X^t + \sigma_n^2 \Phi \right\}^{-1} \end{aligned} \tag{3}$$

where $r_{I\tilde{I}} = \sigma_I^2 + r_{Iv}$, $\sigma_{\tilde{I}}^2 = \sigma_I^2 + \sigma_v^2 + 2r_{Iv}$, and $r_{I|\tilde{I}} = \sigma_I^2 - (r_{I\tilde{I}})^2/\sigma_{\tilde{I}}^2$. The resulting MMSE is equal to σ_e^2 and is given by

$$\sigma_e^2 = \sigma_I^2 - \sigma_I^4 \underline{u}^t X^t \left\{ X \left[\begin{array}{c|c} \sigma_I^2 I_{(N+L+1)\times(N+L+1)} & 0_{(N+L+1)\times L} \\ \hline 0_{L\times(N+L+1)} & r_{I|\tilde{I}} I_{L\times L} \end{array}\right] X^t + \sigma_n^2 \Phi \right\}^{-1} X \underline{u}. \tag{4}$$

Observe that the filters $\underline{\omega}, \underline{b}$ and MSE σ_e^2 are expressed in terms of σ_v^2 and r_{Iv}. We next compute σ_v^2 and r_{Iv} in terms of $\underline{\omega}, \underline{b}$ and σ_e^2, and propose iteratively solving the equations.

To relate σ_v^2 and r_{Iv} to $\underline{\omega}, \underline{b}$ and σ_e^2, we need to additionally model the relationship between $\hat{I}_k$ and I_k. We shall assume that the MDFE output can be well-approximated as $\hat{I}_k = A I_k + \varepsilon_k$ where ε_k is Gaussian and independent of I_k. A similar assumption was used in [7] to argue that removing bias from the MMSE-DFE improves BER (the bias is equal to $(1-A) I_k$). Under this assumption we can compute σ_v^2 and r_{Iv} as follows. For BPSK we have

$$\sigma_v^2 = 4Q\left(\frac{A}{\sigma_\varepsilon}\right), \quad r_{Iv} = -2Q\left(\frac{A}{\sigma_\varepsilon}\right). \tag{5}$$

For 8-PAM we get

$$\begin{aligned} \sigma_v^2 &= \sum_{i=-3}^{3} \sum_{j=-3}^{i} (2i+1-2j) \, Q\left(\frac{(2i+1)\,A-2j}{\sigma_\epsilon}\right), \\ r_{Iv} &= \sum_{i=-3}^{3} \sum_{j=-3}^{i} \frac{-(2i+1)}{2} Q\left(\frac{(2i+1)\,A-2j}{\sigma_\epsilon}\right) \end{aligned} \tag{6}$$

(here $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$). Hence σ_v^2 and r_{Iv} are expressed in terms of A and σ_ε^2. The effective values of A and σ_ε^2 depend on whether we choose to remove the bias. We consider the biased and unbiased cases separately.

3.2.1 The Biased MDFE with ZF Error Propagation Model

The bias will be retained if we directly slice the equalizer output $\widehat{I}_k$ to get $\widetilde{I}_k = dec(\widehat{I}_k)$, as shown in Figure 1a). In this case, we have

$$\begin{aligned} A &= \underline{\omega}^t X \underline{u}, \\ \sigma_\varepsilon^2 &= \sigma_e^2 - (1 - \underline{\omega}^t X \underline{u})^2 \sigma_I^2, \end{aligned} \tag{7}$$

where $\underline{u}^t = \left[0_{1\times(N+L)}, 1, 0_{1\times L}\right]$ (these expressions follow from (1) and the model assumptions). Hence A and σ_ε^2 are expressed in terms of $\underline{\omega}$ and σ_e^2. Hence we have a system of equations (3), (4), (5) (or (6)), and (7) which can be iteratively solved for the filters $\underline{\omega}, \underline{b}$ and MSE σ_e^2 of the *biased MDFE with ZF error propagation model*, starting with the corresponding values for the conventional MMSE DFE.

3.2.2 The Unbiased MDFE with ZF Error Propagation Model

The bias can be removed by slicing a scaled equalizer output $\overline{I}_k = (1/A)\,\widehat{I}_k$ to get $\widetilde{I}_k = dec(\overline{I}_k)$, as shown in Figure 1b). In this case we have

$$\begin{aligned} A &= 1, \\ \sigma_\varepsilon^2 &= \frac{\sigma_e^2 - (1 - \underline{\omega}^t X \underline{u})^2 \sigma_I^2}{(\underline{\omega}^t X \underline{u})^2}. \end{aligned} \tag{8}$$

Hence A and σ_ε^2 are expressed in terms of $\underline{\omega}$ and σ_e^2. Hence we have a system of equations (3), (4), (5) (or (6)), and (8) which can be iteratively solved for the filter weights $\underline{\omega}, \underline{b}$ and MSE σ_e^2 of the *unbiased MDFE with ZF error propagation model*, starting with the corresponding values for the conventional MMSE DFE.

Remark: For BPSK the biased and unbiased MDFE coincide, while for MPAM with $M > 2$ they are different. Considerable simplification can be done for the expressions for σ_v^2 and r_{Iv} in the unbiased case for $M > 2$. Note that the DFE settings derived in [6] are different from both the biased and unbiased MDFE described here. This is explained by the fact that although a similar ZF error propagation model is used, we iteratively compute $\underline{b}, \underline{w}$ and σ_e^2 while [6] computes $\underline{b}, \underline{w}$ and p where $p = \Pr\{\widetilde{I}_k \neq \widehat{I}_k\}$, where the latter procedure requires additional assumptions.

3.3 MMSE Error Propagation Model

We now examine the more complex structure of the detected symbol error statistics for the IIR MMSE-DFE. Assuming correct past decisions, the slicer output $\hat{I}_k$ can be written as $\hat{I}_k = I_k + e_k$, where $e_k = (\gamma_0 - 1)\, I_k + \sum_{j=-\infty}^{-1} \gamma_j I_{k-j} + \sum_{i=-\infty}^{0} \omega_i \xi_{k-i}$ and $\gamma_j =$

$\sum_{i=-\infty}^{0} \omega_i x_{j-i}$. Due to the residual ISI, $\{e_k\}$ is in general a stationary, uncorrelated, zero-mean (but non-Gaussian) sequence with e_k independent of $\{I_n,\, n < k\}$. Since the slicer is memoryless, v_k is a function of I_k and e_k only, and therefore $\{v_k\}$ is also a stationary sequence with v_k independent of $\{I_n,\, n < k\}$. Now the IIR MMSE-DFE can be viewed as an anticausal (and anticausally invertible) whitening filter which whitens $\left\{\xi_k - \frac{\sigma_n^2}{\sigma_I^2} I_k\right\}$ to $\{e_k\}$, as shown in Figure 2b) [7]. Hence e_k is uncorrelated with $\left\{\xi_n - \frac{\sigma_n^2}{\sigma_I^2} I_n, n > k\right\}$. We summarize and strengthen the above in the following way:

1. $\{v_k\}$ is a stationary zero-mean sequence, jointly stationary with $\{I_k\}$ and $\{\xi_k\}$.
2. v_k is independent of $\{I_n, n < k\}$ and $\left\{\xi_n - \frac{\sigma_n^2}{\sigma_I^2} I_n, n > k\right\}$.

Note in particular that the second order statistics satisfy $r_{Iv}(k) = 0,\ \forall k < 0$, .but that $r_{vv}(k) \neq 0$ and $r_{\xi v}(k) \neq 0$ in general.

We next incorporate this model into the design of the DFE, yielding a modified DFE based on an MMSE error propagation model. The modified DFE is again derived using the MMSE criterion, so as to incorporate the statistical properties of the error propagation model. Note that this MMSE error propagation model is more general than the ZF error propagation model in Section 3.1.

3.4 Modified DFE with MMSE Error Propagation Model

Recall that the output of the MDFE is given as in (1). Define $\underline{I}_{k+}^t = \left[\begin{array}{ccc} I_{k+N+L}, & \cdots & , I_k \end{array}\right]$, so that $\underline{I}_k^t = \left[\begin{array}{cc} \underline{I}_{k+}^t & \underline{I}_{k-}^t \end{array}\right]$. For the MMSE error propagation model in Section 3.3, the MSE can be expressed as

$$\begin{aligned} MSE(\underline{\omega}, \underline{b}) &= \sigma_I^2 - 2\sigma_I^2 \underline{\omega}^t X \underline{u} - 2\underline{\omega}^t X \left[R_{II\text{-}} + R_{Iv}\right] \underline{b} + \underline{\omega}^t \left[\sigma_I^2 X X^t + \sigma_n^2 \Phi\right] \underline{\omega} \\ &\quad + \underline{b}^t \left[\sigma_I^2 I_{L\times L} + R_{vv} + R_{I-v} + R_{I-v}^t\right] \underline{b} + 2\underline{r}_{Iv}^t \underline{b} - 2\frac{\sigma_n^2}{\sigma_I^2} \underline{\omega}^t R_{I+v} \underline{b} \end{aligned} \tag{9}$$

where $R_{vv} \equiv E\left\{\underline{v}_k \underline{v}_k^t\right\}$, $\underline{r}_{Iv}^t \equiv E\left\{I_k \underline{v}_k^t\right\}$, $R_{Iv} \equiv E\left\{\underline{I}_k \underline{v}_k^t\right\} = \left[\begin{array}{cc} R_{I+v}^t & R_{I-v}^t \end{array}\right]^t$, $R_{II\text{-}} \equiv E\left\{\underline{I}_k \underline{I}_{k-}^t\right\} = \sigma_I^2 \left[\begin{array}{cc} 0_{(N+L+1)\times L}^t & I_{L\times L} \end{array}\right]^t$, and other quantities are as in Section 3.1. The optimal feedforward and feedback filters are found by minimizing this MSE, which gives

$$\begin{aligned} \underline{b}^t &= \underline{\omega}^t X R_{I\tilde{I}} R_{\tilde{I}\tilde{I}}^{-1} + \frac{\sigma_n^2}{\sigma_I^2} \underline{\omega}^t R_{I+v} R_{\tilde{I}\tilde{I}}^{-1} - \underline{r}_{Iv}^t R_{\tilde{I}\tilde{I}}^{-1} \\ \underline{\omega}^t &= \left[\sigma_I^2 \underline{u}^t X^t - \underline{r}_{Iv}^t P^t\right] \left\{ X R_{I|\tilde{I}} X^t + \sigma_n^2 \Phi - \frac{\sigma_n^2}{\sigma_I^2} \left[R_{I+v} P^t + P R_{I+v}^t\right] \right. \\ &\quad \left. + \frac{\sigma_n^4}{\sigma_I^4} R_{I+v} R_{\tilde{I}\tilde{I}}^{-1} R_{I+v}^t \right\}^{-1} \end{aligned} \tag{10}$$

where
$$P = X R_{I\tilde{I}} R_{\tilde{I}\tilde{I}}^{-1} + \frac{\sigma_n^2}{\sigma_I^2} R_{I+v} R_{\tilde{I}\tilde{I}}^{-1},$$

and $R_{I\tilde{I}} = R_{II\text{-}} + R_{Iv}$, $R_{\tilde{I}\tilde{I}} = \sigma_I^2 I_{L\times L} + R_{vv} + R_{I-v} + R_{I-v}^t$, $R_{I|\tilde{I}} = \sigma_I^2 I_{(N+2L+1)\times(N+2L+1)} - R_{I\tilde{I}} R_{\tilde{I}\tilde{I}}^{-1} R_{I\tilde{I}}^t$. The resulting MMSE is equal to σ_e^2 and is given by

$$\begin{aligned} \sigma_e^2 &= \sigma_I^2 - \underline{r}_{Iv}^t R_{\tilde{I}\tilde{I}}^{-1} \underline{r}_{Iv} - \left[\sigma_I^2 \underline{u}^t X^t - \underline{r}_{Iv}^t P^t\right] \left\{ X R_{I|\tilde{I}} X^t + \sigma_n^2 \Phi - \frac{\sigma_n^2}{\sigma_I^2} \left[R_{I+v} P^t + P R_{I+v}^t\right] \right. \\ &\quad \left. + \frac{\sigma_n^4}{\sigma_I^4} R_{I+v} R_{\tilde{I}\tilde{I}}^{-1} R_{I+v}^t \right\}^{-1} \left[\sigma_I^2 X \underline{u} - P \underline{r}_{Iv}\right]. \end{aligned} \tag{11}$$

Note that if $\underline{r}_{Iv}$ and R_{I+v} are zero (as are implied by the ZF error propagation model), then the MMSE error propagation model reduces to the ZF error propagation model. Now observe that the filters $\underline{\omega}, \underline{b}$ and MSE σ_e^2 are expressed in terms of R_{vv} and R_{Iv}, whose components depend on $r_{vv}(m)$, for $m = 0, \pm 1, ..., \pm(L-1)$, and $r_{Iv}(m)$, for $m = (1-L), ..., (N+L+1)$. We next compute these $r_{vv}(m)$ and $r_{Iv}(m)$ (and hence R_{vv} and R_{Iv}) in terms of $\underline{\omega}, \underline{b}$ and σ_e^2, and propose iteratively solving the equations.

To relate $r_{vv}(m)$ and $r_{Iv}(m)$ to $\underline{\omega}, \underline{b}$ and σ_e^2, we shall further assume that the MDFE output can be well-approximated as $\hat{I}_k = AI_k + \varepsilon_k$, where ε_k, ε_{k-m} are conditionally Gaussian and independent given I_k (for $m > 0$), and $\{\varepsilon_n, I_{n+1}, n \geq k\}$ is independent of $\{I_n, n \leq k\}$. Related and similarly justified assumptions were used in Section 3.2 and in [7]. Under this assumption we can compute $r_{vv}(m)$ and $r_{Iv}(m)$ as follows. For BPSK we can write

$$\begin{aligned} r_{vv}(m) &= \begin{cases} 2Q\left(\frac{A}{\sigma_\varepsilon}\right)\left[Q\left(\frac{A+A_{|m|}}{\sigma_{\varepsilon|I}(m)}\right) - Q\left(\frac{A-A_{|m|}}{\sigma_{\varepsilon|I}(m)}\right)\right], & m \neq 0 \\ 4Q\left(\frac{A}{\sigma_\varepsilon}\right), & m = 0 \end{cases} \\ r_{Iv}(m) &= \begin{cases} Q\left(\frac{A-A_m}{\sigma_{\varepsilon|I}(m)}\right) - Q\left(\frac{A+A_m}{\sigma_{\varepsilon|I}(m)}\right), & m > 0 \\ -2Q\left(\frac{A}{\sigma_\varepsilon}\right), & m = 0 \\ 0, & m < 0 \end{cases} \cdot \end{aligned} \tag{12}$$

For 8-PAM we can get similar results which are too lengthy to be included here. Hence $r_{vv}(m)$ and $r_{Iv}(m)$ are expressed in terms of A, $A_m, \sigma_\varepsilon^2$ and $\sigma_{\varepsilon|I}^2(m)$. The effective values of A, $A_m, \sigma_\varepsilon^2$ and $\sigma_{\varepsilon|I}^2(m)$ depend on whether we choose to remove the bias. We considered the biased and unbiased cases separately.

3.4.1 The Biased MDFE with MMSE Error Propagation Model

The bias is retained if we directly slice the equalizer output $\hat{I}_k$ as shown in Figure 1a). In this case, we have

$$\begin{aligned} A &= \underline{\omega}^t X \underline{u}_0; \quad A_m = \underline{\omega}^t X \underline{u}_m, \\ \sigma_\varepsilon^2 &= \sigma_e^2 - \left(1 - \underline{\omega}^t X \underline{u}_0\right)^2 \sigma_I^2; \quad \sigma_{\varepsilon|I}^2(m) = \sigma_e^2 - \left(\underline{\omega}^t X \underline{u}_0 - 1\right)^2 \sigma_I^2 - \left(\underline{\omega}^t X \underline{u}_m\right)^2 \sigma_I^2. \end{aligned} \tag{13}$$

where $\underline{u}_0^t = \left[0_{1\times(N+L)}, 1, 0_{1\times L}\right]^t$ and $\underline{u}_m^t = \left[0_{1\times(N+L-m)}, 1, 0_{1\times(L+m)}\right]^t$ (these expressions follow from (1) and the model assumptions). Hence A, $A_m, \sigma_\varepsilon^2$ and $\sigma_{\varepsilon|I}^2(m)$ are expressed in terms of $\underline{\omega}$ and σ_e^2. Hence we have a system of equations (10), (11), (12), and (13) which can be iteratively solved for the filters $\underline{\omega}, \underline{b}$ and MSE σ_e^2 of the *biased MDFE with MMSE error propagation model*, starting with the corresponding values for the conventional MMSE DFE.

3.4.2 The Unbiased MDFE with MMSE Error Propagation Model

The bias can be removed by slicing a scaled equalizer output $\overline{I}_k = (1/A)\hat{I}_k$ as shown in Figure 1b). In this case we have

$$\begin{aligned} A &= 1; \quad A_m = \frac{\underline{\omega}^t X \underline{u}_m}{\underline{\omega}^t X \underline{u}_0}, \\ \sigma_\varepsilon^2 &= \frac{\sigma_e^2 - \left(1 - \underline{\omega}^t X \underline{u}_0\right)^2 \sigma_I^2}{\left(\underline{\omega}^t X \underline{u}_0\right)^2}; \quad \sigma_{\varepsilon|I}^2(m) = \frac{\sigma_e^2 - \left(\underline{\omega}^t X \underline{u}_0 - 1\right)^2 \sigma_I^2 - \left(\underline{\omega}^t X \underline{u}_m\right)^2 \sigma_I^2}{\left(\underline{\omega}^t X \underline{u}_0\right)^2}, \end{aligned} \tag{14}$$

Hence A, $A_m, \sigma_\varepsilon^2$ and $\sigma_{\varepsilon|I}^2(m)$ are expressed in terms of $\underline{\omega}$ and σ_e^2. Hence we have a system of equations (10), (11), (12), and (14) which can be iteratively solved for the filters $\underline{\omega}, \underline{b}$ and MSE σ_e^2 of the *unbiased MDFE with MMSE error propagation model*, starting with the corresponding values for the conventional MMSE DFE.

4 Performance Analysis

The standard union bound on the BER of ML soft-decision decoding of a convolutional code over an AWGN channel does not account for error propagation in the DFE. However, it is straightforward to modify the union bound using the stationary error models we have introduced, treating the error term as additional white noise (ZF error propagation model) or correlated noise (MMSE error propagation model). Note that in the latter case, the Viterbi algorithm is mismatched, i.e., it assumes white noise. Although based on a crude approximation at best to the true Markov error propagation model, such a bound may give some useful analytical measure of performance degradation due to error propagation in the convolutionally coded system. The modified union bound is [9]

$$P_e \leq \frac{1}{k}\sum_{d=d_{min}}^{\infty}\sum_{m=1}^{A_d} w\left(S_d^{(m)}\right) Q\left(\frac{\left\|S_d^{(m)}-S_0\right\|^2}{2\sqrt{\left(S_d^{(m)}-S_0\right)^t R_{ee}\left(S_d^{(m)}-S_0\right)}}\right)$$

where k/n is the rate of the convolutional code, A_d is the number of codewords with weight d, $S_d^{(m)}$ is the $m-$th codeword with weight d, $w\left(S_d^{(m)}\right)$ is the Hamming weight of the information sequence corresponding to $S_d^{(m)}$, and R_{ee} is the correlation matrix for the equalizer output noise. For the ZF error propagation model, $R_{ee} = \sigma_e^2 I$ can be obtained by substituting the DFE or MDFE filter taps in (2), the expression for σ_e^2. For the MMSE error propagation model, R_{ee} (which is not diagonal) can be similarly obtained by substituting the DFE or MDFE filter taps in an expression for $r_{ee}(m)$ (not shown).

5 Simulation Results

Simulations using BPSK and 8-PAM were carried out on a typical telephone channel to evaluate the performance of the MDFE. The channel has the following sampled impulse: $[x_{-7}, ..., x_7] = [-0.1, -0.1,\ -0.15,\ -0.3, -0.33,\ -0.16, 0.33, 1,\ 0.33,\ 0.16,\ -0.33,\ -0.3, -0.15, -0.1, -0.1]$. It has a deep spectral null at DC which results in severe ISI. We used 22 feedforward and 28 feedback filter taps with BPSK, and 50 feedforward and 28 feedback taps with 8 PAM. Due to the limited space, we present primarily BPSK results, with one additional 8-PAM result to illustrate the importance of bias removal for the MDFE with convolutional coding (recall that with BPSK the biased and unbiased MDFE coincide).

In Figure 3a), we show the BER for the MDFE and the DFE with BPSK and no coding. In Figure 3b) we plot the histograms for the burst errors lengths at SNR=8 and 16 dB. It is seen that at the lower SNR the number of isolated errors for the MDFE is about the same as for the DFE, while at the higher SNR the number of isolated errors for the MDFE is larger than for the DFE. For all other burst lengths, the MDFE has fewer errors than the DFE. Evidently, the MDFE replaces longer burst errors with either isolated errors or shorter burst errors. This does not translate into improved BER for the following reason: at low to moderate SNR the increase in isolated errors offsets the decrease in propagated errors for the MDFE relative to DFE, while at high SNR the isolated errors dominate performance and are unchanged for the MDFE relative to DFE. Hence, the performance of the MDFE and DFE without coding is similar, as seen in Figure 3a). Furthermore, nearly identical performance was obtained for the MDFE with ZF versus MMSE error propagation models (only the prior is plotted).

Next we examine how the MDFE's ability to spread out burst errors can be advantageous when combined with suitable coding schemes (a convolutional code in this paper). The parameters for the convolutional codes in our simulations are listed in Table 1. The decoding delay was set equal to $5K$, where K is the constraint length. In Figure 4 we show the BER for the MDFEs and DFE with BPSK and the various convolutional

Table 1: Convolutional code parameters

Rate	K	Generator Matrix	d_{free}	Rate	K	Generator Matrix	d_{free}
2/3	2	17 6 15	3	1/4	4	13 15 15 17	13
1/3	3	5 7 7	8	1/5	3	7 7 7 5 5	13

codes. The results show that the MDFE can provide as much as 4-6 dB performance gain at low SNRs over the conventional DFE with coding. Note also that the MDFE with MMSE error propagation model shows only a marginal performance gain over the MDFE with ZF error propagation model. In Figure 5 we compare the simulated BER curves with the bounds in Section 4 for BPSK and the rate 1/3 and 2/3 convolutional codes. In Figure 6 we show the symbol error rate (SER) for the biased and unbiased MDFEs and DFEs with 8-PAM and the rate 1/3 convolutional code. Note the improved performance with the unbiased MDFE compared to the biased MDFE when combined with the convolutional code; by contrast, nearly identical performance was obtained with the unbiased and biased DFE (only the prior is plotted). In Figure 7 we show the BER for the MDFE and DFE with BPSK and the rate 1/4 convolutional code, and (for the DFE) a convolutional interleaver of various depths. The DFE with interleaving does not show much improvement at low-to-moderate SNRs, since the interleaver cannot effectively randomize the burst errors, whereas the MDFE without interleaving can provide 5-6 dB improvement. Even at moderate SNRs where the DFE with interleaving performs about the same as the MDFE without interleaving, the MDFE is still a better choice than the DFE since the interleaver introduces a large amount of delay. For example, for the DFE with interleaving to achieve the same BER of 2×10^{-3} at SNR=14 dB as the MDFE without interleaving, an interleaving delay of about 180 symbol intervals corresponding to a convolutional interleaver of depth 20 is required. Note that the total system delay for the MDFE without interleaver is the sum of the delays introduced by the matched filter, feedforward filter, and Viterbi decoder, and is equal to $14 + 22 + 5 * 4 = 56$, while the total delay for the DFE with interleaver equals $56 + 180 = 236$ to achieve comparable performance.

6 Summary

We proposed two stationary error models for the DFE, and derived corresponding modified DFEs (MDFEs), which differ from the conventional DFE only in their tap settings. The MDFE tap settings attempt to jointly compensate for estimator bias and error propagation. The MDFEs can effectively reduce burst error lengths and hence achieve large performance gain over the DFE when each is combined with the same convolutional code (the performance of the MDFE and DFE in uncoded systems is essentially the same).To achieve the same performance, the DFE additionally requires an interleaver, which greatly increases the delay. In this sense, the MDFE is optimized for the convolutional code, and can be viewed as a joint equalization and interleaving technique.

References

[1] J. G. Proakis. *Digital Communications.* McGraw-Hill, 3rd edition, 1995.

[2] H. Meyr, M. Moeneclaey, and S. A. Fechtel. *Digital Communications Receivers: Synchronization, Channel Estimation and Signal Processing.* Wiley, 1998.

[3] D. L. Duttweiler, J. E. Mazo, and D. G. Messerschmitt. An upper bound on the error probability in decision-feedback equalization. *IEEE Transactions on Information Theory*, IT-20(4):490–497, July 1974.

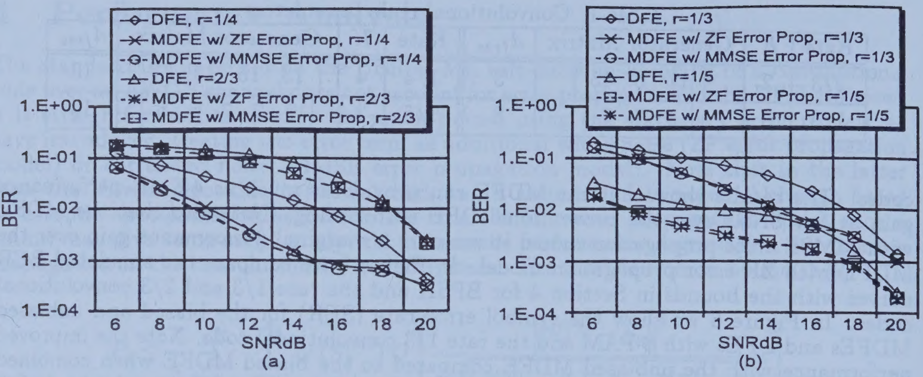


Figure 4. BER curves for the MDFEs and DFE with BPSK and convolutional code. The code rates are (a)1/4, and 2/3, and (b)1/3, and 1/5.

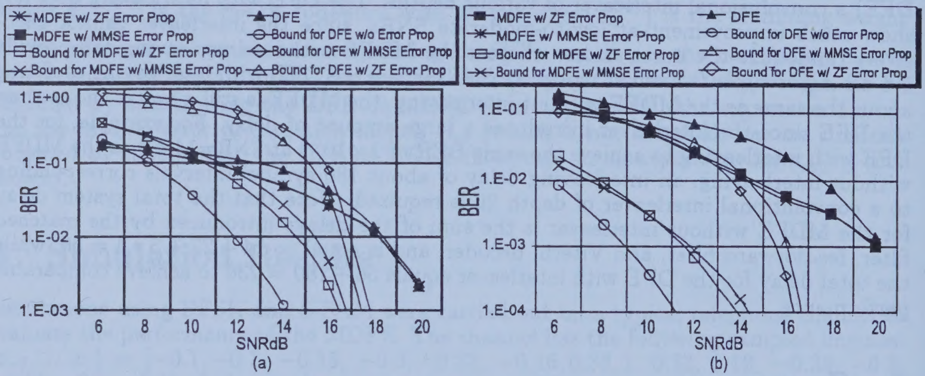
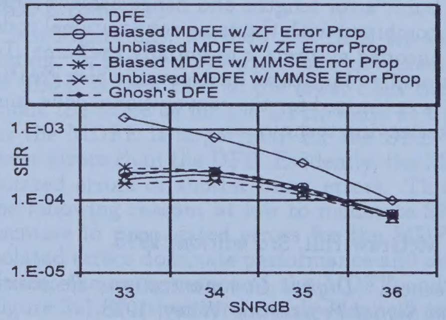


Figure 5. Bounds on BER for DFEs and MDFEs by modeling the burst errors with either colored noise(MMSE model) or white noise(ZF model) with enhanced noise variance and convolutional code. The code rate is (a) 1/3, (b) 2/3.



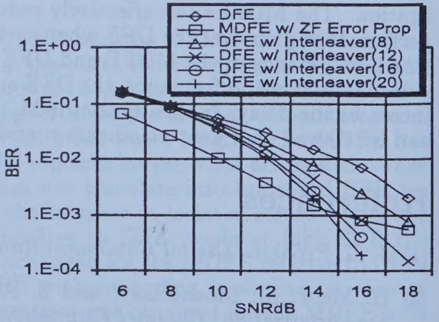


Figure 6. SER curves for DFE and MDFE with 8PAM and convolutional coding. The code rate is 1/3.

Figure 7. BER curves for ZF-MDFE, DFE and DFE with convolutional interleaver of different depths (in parenthesis), with BPSK and convolutional coding. The code rate is 1/4.

[4] M. Chiani. Introducing erasures in decision-feedback equalization to reduce error propagation. *IEEE Transactions on Communications*, 45(7):757–760, July 1997.

[5] J. W. M. Bergmans, J. O. Voorman, and H. W. Wong-Lam. Dual decision feedback equalizer. *IEEE Transactions on Communications*, 45(5):514–518, May 1997.

[6] M. Ghosh. Analysis of the MMSE-DFE with error propagation. *IEEE Global Telecommunications Conference*, pages 85–89, November 1997.

[7] J. Cioffi, G. Dudevoir, M. Eyuboglu, and G.D. Forney. Minimum mean-square-error decision feedback equalization and coding —part i: Equalization results. *IEEE Transactions on Communications*, 43(10):2582–2594, October 1995.

[8] N. Al-Dhahir and J. M. Cioffi. MMSE decision-feedback equalizers: Finite-length results. *IEEE Transactions on Information Theory*, 41(4):961–975, July 1995.

[9] O. E. Agazzi and N. Seshadri. On the use of tentative decisions to cancel intersymbol interference and nonlinear distortion (with application to magnetic recording channels). *IEEE Transactions on Information Theory*, 43(2):394–408, March 1997.

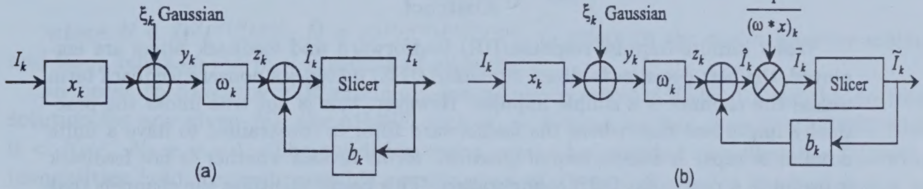

ξ_k Gaussian
I_k
x_k
y_k
ω_k
z_k
$\hat{I}_k$
Slicer
$\tilde{I}_k$
b_k
(a)
ξ_k Gaussian
$\frac{1}{(\omega * x)_k}$
I_k
x_k
y_k
ω_k
z_k
$\hat{I}_k$
$\bar{I}_k$
Slicer
$\tilde{I}_k$
b_k
(b)

Figure 1. The Equivalent Discrete-time Channel Model with (a) the DFE, (b) the unbiased DFE receiver.

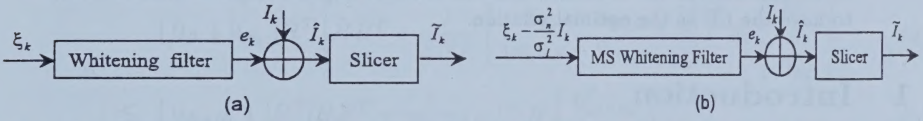

ξ_k
Whitening filter
e_k
I_k
$\hat{I}_k$
Slicer
$\tilde{I}_k$
(a)
$\xi_k - \frac{\sigma_n^2}{\sigma_I^2} I_k$
MS Whitening Filter
e_k
I_k
$\hat{I}_k$
Slicer
$\tilde{I}_k$
(b)

Figure 2. Whitening Filter Representation of (a) ZF-DFE, (b) MMSE-DFE.

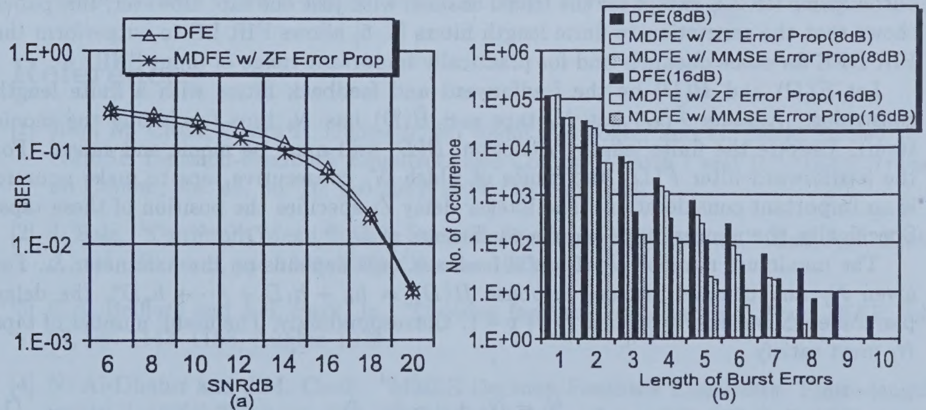


Figure 3. (a) BER curves, (b) burst error lengths at SNR=8, 16dB for DFE and MDFE on a typical telephone channel with uncoded BPSK.

When the Best Decision-Feedback Equalizer is a Linear Equalizer[*]

Wei Shi and Rick Wesel
Electrical Engineering Department
University of California, Los Angeles
Los Angeles, CA 90095-1594
wshi@ee.ucla.edu and wesel@ee.ucla.edu

Abstract

When infinite-impulse-response(IIR) feedforward and feedback filters are employed, the best decision-feedback equalizer(DFE) includes a nonzero feedback term unless the channel is a simple impulse. However, this is not true under the practically important case where the feedforward filter is constrained to have a finite number of taps. A comparison of quadratic forms reveals whether or not feedback is useful in a particular DFE configuration. This paper identifies the channels that perform better with finite-impulse-response (FIR) linear equalizer(LE) than all the possible FIR DFE for a given number of feedforward filter taps and a specific input SNR, and derives the condition that the channel impulse response should satisfy to have the LE as the optimal solution.

1 Introduction

Most analysis of DFE concerns the DFE with IIR filters [1, 2, 3]. IIR DFE strictly outperforms IIR LE except for the trivial channel with just one tap. However, this paper shows that the restriction to finite length filters [4, 5] allows FIR LE to outperform the FIR DFE for some channels and for practically important range of input SNR.

Let $\tilde{F}(D)$ and $\tilde{B}(D)$ be the feedforward and feedback filters with a finite length constraint. Assume $\tilde{F}(D)$ has N_f taps and $\tilde{B}(D)$ has N_b taps (excluding the monic term). Despite the finite length restriction, $\tilde{B}(D)$ still must be monic and causal. For the feedforward filter $\tilde{F}(D)$, the choice of which N_f consecutive taps to make nonzero is an important consideration. The integer delay $\triangle$ specifies the position of these taps. Specifically, the nonzero taps are the coefficients of $D^{-\triangle}, \cdots, D^{N_f-\triangle-1}$.

The maximum number N_b of useful feedback taps depends on the parameter $\triangle$. For given N_f and channel impulse response $H(D) = h_0 + h_1 D + \cdots + h_v D^v$, the delay parameter $\triangle$ can vary from 0 to $N_f + v - 1$. Correspondingly, The useful number of taps N_b must satisfy

$$N_b = N_f + v - 1 - \triangle \tag{1}$$

Note that, when $N_b = 0$, the FIR DFE is actually a FIR LE.

---

[*] This research is supported by NSF CAREER award #9733089.

2 Channel conditions for FIR LE to be optimal

This section identifies what kind of channels can perform better with FIR LE($N_b = 0$) than all the possible FIR DFEs ($0 < N_b \leq N_f + v - 1$) for some given N_f and a specific input SNR.

Let the channel impulse response $H(D) = h_0 + h_1 D + \cdots + h_v D^v$. Define two matrices H and $\tilde{H}$ as

$$H = \begin{bmatrix} h_0 & h_1 & \cdots & h_v & 0 & \cdots & 0 \\ 0 & h_0 & h_1 & \cdots & h_v & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_0 & h_1 & \cdots & h_v \end{bmatrix} \tag{2}$$

and

$$\tilde{H} = \begin{bmatrix} \multicolumn{2}{c}{H} \\ 0_{N_b \times (\Delta+1)} & I_{N_b} \end{bmatrix} \tag{3}$$

where $H \in R^{N_f \times (N_f+v)}$, $\tilde{H} \in R^{(N_f+N_b)\times(N_f+v)}$, I_n refers to the $n \times n$ identity matrix and $0_{m\times n}$ refers to the $m \times n$ matrix of all zeros.

In order to have the FIR minimum-mean-square error(MMSE) LE as the optimal solution for any given N_f, the MMSE with $N_b = 0$ must be less than the MMSE with $0 < N_b \leq N_f + v - 1$. The condition turns out to be satisfied provided the following inequalities hold for each possible combination of Δ and N_b under the restriction of $\Delta + N_b = N_f + v - 1$ and $0 < N_b \leq N_f + v - 1$.

$$[\, 0_\Delta \; 1 \; 0_{N_b} \,]\tilde{H}^T \left[\tilde{H}\tilde{H}^T + \frac{1}{SNR} \begin{bmatrix} I_{N_f} & 0_{N_f \times N_b} \\ 0_{N_b \times N_f} & 0_{N_b \times N_b} \end{bmatrix} \right]^{-1} \tilde{H} \begin{bmatrix} 0_\Delta^T \\ 1 \\ 0_{N_b}^T \end{bmatrix} \tag{4}$$

$$\leq \; [\, 0_{\Delta+N_b} \; 1 \,] H^T [H H^T + \frac{1}{SNR} I_{N_f}]^{-1} H \begin{bmatrix} 0_{\Delta+N_b}^T \\ 1 \end{bmatrix} \tag{5}$$

where 0_n refers to the n-element row of all zeros. Here $SNR = \frac{\bar{\mathcal{E}}_w}{N_0/2}$ is the input SNR.

References

[1] John M. Cioffi, Glenn P. Dudevoir, M. Vedat Eyuboğlu, and G. David Forney, Jr., "MMSE Decision-Feedback Equalizers and Coding – Parts I and II," *IEEE Trans. on Comm.*, vol. 43, no. 10, pp. 2582–2604, October 1995.

[2] J. Salz, "Optimum Mean Square Decision Feedback Equalization," *Bell Syst. Tech. J.*, vol. 52, no. 8, pp. 1341–1373, October 1973.

[3] C.A. Belfiore and J.H. Park, Jr., "Decision feedback equalization," *Proc. IEEE*, vol. 67, pp. 1143–1156, August 1979.

[4] N. Al-Dhahir and J.M. Cioffi, "MMSE Decision-Feedback Equalizers: Finite-length results," *IEEE Trans. on Info. Theory*, vol. 41, no. 4, pp. 961–975, July 1995.

[5] John M. Cioffi, "Signals and Detection," Stanford University Lecture Notes, 1994.

PSEUDO-ORTHOGONAL BASES FOR MULTICARRIER MODULATION

*Xie Yongzhe *   Vellenki U. Reddy #   and   Ko Chi Chung **
** Department of Electrical Engineering, National University of Singapore,*
Singapore 119260,   Email: {yzxie,elekocc}@eeserver.ee.nus.edu.sg
# Department of Electrical Communication Engineering, India Institute of Science,
Banglore 560012, India   Email: vur@ece.iisc.ernet.in

INTRODUCTION

In multicarrier modulation, a transmission channel is partitioned into a number of sub-channels, each with its own associated carrier. All of the current available multicarrier systems employ orthogonal carriers. Discrete multitone modulation (DMT), or orthogonal frequency division multiplexing (OFDM), uses DFT bases, while overlapped discrete multitone or discrete wavelet multitone modulation (DWMT) [1] uses orthogonal wavelet bases which have better spectral containment compared to the DFT bases, thereby resulting in a higher level of immunity to channel noises and inter-channel interference (ICI). However, this advantage is lost to some extent because the bases in the DWMT extend much beyond the block length, thereby suffering more from inter-block interference (IBI). In addition, longer bases lead to more system complexity (i.e. the complexity of the modulator, the demodulator and the equalizer) and higher system delay [2]. Therefore, the ideal bases in a multicarrier modulation system for severe interference plus noise channel should be of less overlap between blocks in the time domain and have a high stop-band attenuation in the frequency domain. However, bases satisfying these two conditions and the orthogonality while having fast algorithm to implement are hard to find.

In this paper, we propose a DWMT system with pseudo-orthogonal bases. We first examine the justification for replacing the orthogonal bases with pseudo-orthogonal bases in DWMT system and then compare the performance of DMT and DWMT with two different bases (i.e. orthogonal and pseudo-orthogonal). We may note here that the orthogonal wavelet bases and pseudo-orthogonal bases discussed in this paper are designed by cosine modulation, which can be implemented by fast algorithm [3].

WHY PSEUDO-ORTHOGONAL BASES

The basic multicarrier modulation system is given in Figure 1. In DMT, the synthesis filters { $F_k(z)$ } in the transmitter correspond to DFT bases, whose length L is equal to the number of sub-channels M. In DWMT, { $F_k(z)$ } with length $L = gM$, form a set of perfect reconstruction paraunitary filter bank whose impulse responses constitute the M-band wavelet bases, and g is the overlap factor. The impulse responses { $f_k(n)$ } of the synthesis filters { $F_k(z)$ } satisfy:

$$\sum_n f_k(n) f_l(n+rM) = \delta(k-l)\delta(r) \qquad 0 \le k,l \le M-1 \tag{1}$$

The impulse response of the k^{th} analysis filter $F_k'(z)$ in the receiver, is related to $f_k(n)$ as

$$f_k'(n) = f_k^*(d+L-n) \tag{2}$$

where $f_k^*(n)$ denotes complex conjugate of $f_k(n)$, and d is such that $d + D = tM$, with D and t denoting the channel delay and some non-negative integer, respectively.

Practical channel will introduce a loss of orthogonity among the bases, thus causing two kinds of inter-symbol interference (ISI), inter-channel interference (ICI) among signals of different sub-channels and inter-block interference (IBI) among signals belong to different blocks. To mitigate the effect of ISI, a post-detection equalizer is used which combines the receiver output samples to give

$$\hat{s}_{m_1}(i_1) = \sum_{k=-v}^{v} \sum_{m' \in \Omega(m_1)} \lambda_{m_1}(m',k)\theta_{m'}(i_1 + k + g_1) \tag{3}$$

Here $\hat{s}_{m_1}(i_1)$ is an estimate of $s_{m_1}(i_1)$, $\Omega(m_1)$ denotes the set of indices of the sub-channels that are used in the estimation of the symbols in the m_1^{th} sub-channel, v denotes the number of blocks prior to and after the desired block, and g_1 is the symbol delay for each sub-channel with $g_1 = t + g$.

The design of the equalizer is equivalent to choosing the parameters $\lambda_{m_1}(m',k)$. If the overlap factor g is large, a higher value of v is needed. On the other hand, if there is more ICI, more neighboring sub-channels of the m_1^{th} sub-channel are to be included in $\Omega(m_1)$.

If $F_k(z)$ has very high stopband attenuation for all k, then a relatively small number of neighboring sub-channels will have significant spectral overlap with $F_k(z)$. Under these conditions, ICI will become very small thereby reducing the number of sub-channels to be combined in post-detection equalizer. Usually, however, we have to increase the length of the orthogonal bases to get higher stopband attention, which will lead to more overlap in time between different blocks, i.e. more IBI.

The orthogonality of the bases will be lost after the carriers pass through the non-ideal channel and an equalizer has to be used anyway. We, therefore, can relax the orthogonality requirement of the bases to some extent and lay more emphasis on minimizing the stopband energy. The error caused by the loss of orthogonality of the bases together with the channel distortion can be compensated in the post-detection equalizer. Pseudo-orthogonal bases offer a greater degree of freedom in their design compared to the orthogonal bases for same L (length of the prototype filter) and M, and therefore, they can be designed with higher stopband attenuation without increasing the length. We can also allow relatively higher loss in orthogonality of the adjacent sub-channels to further increase the stopband attenuation. As a result, the errors in a sub-channel caused by ISI and non-orthogonality will be mainly contributed from the adjacent sub-channels. This means, we only need to combine the outputs from the desired sub-channel and its adjacent sub-channels in the equalizer.

SIMULATION RESULTS

We conducted simulations to compare the performance of DMT and DWMT with orthogonal and pseudo-orthogonal bases. In all the cases, we used $M = 16$ and 16-QAM signal constellation in each sub-channel. The complex pseudo-orthogonal bases were designed according to [4][5]. The impulse response of channel $C(z)$, used in the simulations, was given by [-0.12 -0.13 -0.16 -0.18 -0.22 -0.26 -0.12 0.68 0.46 0.26 0.07 -0.04 -0.1 -0.12]. The post-detection equalizer was designed according to [1] to achieve the maximum Signal to Interference and Noise Ratio. We chose $v = 1$ in the DMT case and $v = g$ for DWMT. $|\Omega(m_1)|$ denotes the size of $\Omega(m_1)$. If $|\Omega(m_1)| = 6$, then $\Omega(m_1)$ includes three conjugate pairs of sub-channels: $(m_1 - 1,\ M - m_1)$ $(m_1,\ M - m_1 - 1)$, $(m_1 + 1,\ M - m_1 - 2)$. The channel impulse response was estimated using a 1000-bit training sequence with channel noise present. This was repeated 100 times with different channel noise realizations and the average bit-error rate was computed. The noise was assumed to be a zero mean white

Gaussian process. Table 1 gives the number of bits in error for different schemes. We may point out that the training bits are not counted in calculating the average bit error probability.

Note from Table 1 that the bit-error rate is high in the zeroth sub-channel in all the cases. This is because the channel has a deep spectral null in the region covered by the zeroth sub-channel. DWMT technique with both orthogonal and pseudo-orthogonal bases performs much better than the DMT due to the higher spectral containment of its bases, and hence, less affected by ICI. When $|\Omega(m_1)| = 6$, the DWMT with pseudo-orthogonal bases performs better than the DWMT with orthogonal bases of the same length because of the higher stopband attenuation that can be achieved with pseudo-orthogonal bases. To match the performance with that of pseudo-orthogonal bases, we have to increase $|\Omega(m_1)|$ from 6 to 16, i.e., we have to include all the sub-channels in the post-detection equalizer in the case of orthogonal bases, thereby increasing the computation complexity of the equalizer.

REFERENCES

[1] S. D. Sandberg and M. A. Tzannes, “Overlapped Discrete Multitone Modulation for high Speed Copper Wire Communications”, IEEE Journal On Selected. Areas In Communication, Vol. 13, No. 9, Dec. 1995.

[2] G. W. Wornell, "Emerging Applications of Multirate Signal Processing and Wavelets in Digital communications", Proceedings of the IEEE, Vol 84, No 4, Apr. 1996.

[3] P. P. Vaidyanthan, Multirate Systems and Filter Banks, Englewood Cliffs, NJ: Prentice Hall, Inc., 1993.

[4] Tushar K. Adhikary and Vellenki U. Reddy, “Complex wavelet packets for multicarrier modulation”, pp.1821-1824, Proceedings of 1998 IEEE International Conference on Acoustics, Speech and Signal Processing

[5] Xie Yongzhe "A Novel Trellis-Coded Multicarrier Modulation System", Master's Thesis, Department of Electrical Engineering, the National University of Singapore

Table 1 Number of bits in error per 6000 bits in each sub-channel (per 3000 bits in sub-channels 0&8 for DMT) for different bases with 16-QAM constellation (SNR=30dB)

Type of bases	$\lvert\Omega(m_1)\rvert$	v	Sub-channel number									Average bit-error probability
			0	1	2	3	4	5	6	7		
Pseudo(2),L=32	6	2	909	1	1	8	8	15	9	19		0.0202
Orthogonal Bases, L=32	6	2	968	2	11	71	120	222	161	241		0.0374
	16	2	881	1	2	8	6	15	9	23		0.0197
DMT	$\lvert\Omega(m_1)\rvert$	1	Sub-channel number									Average bit-error probability
			0	1	2	3	4	5	6	7	8	
	16		864	29	10	43	93	240	327	406	378	0.0498

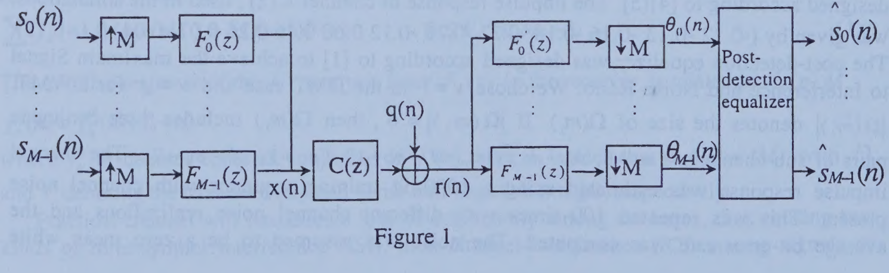


Figure 1

Maximum-Likelihood Carrier Frequency Offset Estimation in Orthogonal Frequency Division Multiplexing(OFDM) Systems

Ayesha T. Huq, Erdal Panayırcı* and Costas N. Georghiades
Department of Electrical Engineering
Texas A&M University,
College Station, TX 77843-3128
*corresponding author e-mail: *eepanay@ee.tamu.edu*

Abstract: In this paper a Maximum Likelihood(ML) estimation algorithm is derived for frequency offset estimation in OFDM systems with M-PSK modulation. Its mean-squared performance is obtained analytically and compared with simulation results. The algorithm exploits the redundant information contained in the cyclic prefix preceeding the OFDM symbols, thus reducing the need for pilots. It is observed that the resulting algorithm generates very accurate estimation even when the offset is high. It is also shown that the frequency estimator may be used in a tracking mode.

1 Introduction

OFDM signaling is proven to be an effective way to overcome the effects of fading channel and multi-path by dividing the frequency selective fading channel into a number of sub-channels corresponding to the OFDM sub-carrier frequencies. See [1] and [2] and the references therein. When compared with single carrier systems, channel equalization is less complex, and sensitivity to channel estimation and frame synchronization error can be reduced. However, OFDM systems are more sensitive to carrier frequency offset, caused by the mismatch of the oscillators in the transmitter and the receiver, than single carrier systems are. This sensitivity to frequency offsets is pointed out to be one of the major OFDM disadvantages. A time varying frequency error not only disturbs the frequency orthogonality, but also makes sub-carrier synchronization much more difficult to achieve and maintain [3],[4]. So, estimating the frequency offset of OFDM systems has become a challenging area of interest in recent years.

A frequency offset estimate may be generated at the receiver with the aid of pilot symbols known at the receiver [5], [6] or as in [7], by maximizing the average log-likelihood function. Redundancy in the transmmited OFDM signal also offers the opportunity for synchronization. Such an approach is found in [8], [9] and [10] for a frequency offsett. Along with this approach, Van de Beek [8] describes a method of using a correlation with the cyclic prefix to find the frequency offset and the OFDM frame timing. However they make a Gaussian approximation for data in deriving the maximum likelihood ratio and therefore the effect of modulation of data symbols is ignored.

This paper extends the Van de Beek's method in several ways as follows: i) A maximum likelihood estimation algorithm is derived for frequency offset estimation in OFDM systems without any Gaussian assumption of data symbols. ii) Its mean-squared performance is obtained analytically and compared with simulation results. iii) it is show that the derived estimator can work also in tracking mode as well as the forward estimation mode. iv) A tracking algorithm is provided to track the frequency offset and the symbol error rate(SER) degradation before and after synchronization with estimated offset is also calculated for a closed loop system. We conclude that estimator performance improves as SNR increases but cyclic prefix length plays a key role in controlling the performance. Also a closed loop system realization shows that performance is improved with the suggested estimator.

2 OFDM Systems

The main idea behind OFDM is to split the data stream to be transmitted into N parallel streams of reduced data rate and to transmit each of them on a separate sub-carrier. These carriers are made orthogonal by appropriately choosing the frequency spacing between them. Therefore, spectral overlapping among the sub-carriers is allowed, since the orthogonality will ensure that the receiver can separate the OFDM sub-carriers, and a better spectral efficiency can be achieved than by using simple frequency division multiplex. A typical block diagram of an OFDM system is the one depicted in [8]. During each ith OFDM symbol interval, the transmitted data symbols split by a serial-to-parallel conversion block into N lower bit rate streams $\{a_i^{k-1}\}_{k=1}^{N}$ each modulating a sub-carrier. These sub-carriers are converted via an inverse FFT processor into N time signals $\{s_i^{k-1}\}_{k=1}^{N}$, and then a cyclic prefix of length L is added to form a complete OFDM symbol. The cyclic prefix is generated by copying the last L samples of the OFDM symbol (N samples long) and appending them to the beginning of the frame. Therefore, the effective length of the OFDM symbol as transmitted is $N+L$ samples long. The insertion of a cyclic prefix is an accepted means of avoiding intersymbol interference and preserving orthogonality between sub-carriers. It will be shown that the redundant information contained within the cyclic prefix enables also the frequency offset estimation very effectively and without additional pilo ts.

We assume that the N complex data being transmitted is derived from M-PSK (M-ary phase shift keying) scheme. The resulting OFDM symbol denoted by $s(k)$ is serially transmitted over a discrete-time channel, whose impulse response is assumed to be shorter than L samples. At the receiver the data are recovered by means of a DFT (Discrete Fourier Transform). Assuming that the OFDM frame and symbol timing synchronization have already been achieved and that the OFDM signal is transmitted over an AWGN channel, the complex signal envelope of the received signal can be expressed as

$$r(k) = s(k)\exp\left[j(\frac{2\pi\epsilon k}{N} + \phi)\right] + n(k) \tag{1}$$

where, ϵ is the relative frequency offset of the channel (the ratio of the actual frequency offset to the intercarrier spacing), ϕ is the channel phase offset and $n(k)$ is the complex envelope of the additive white Gaussian noise with variance $\sigma_n^2 = E\{|n(k)|^2\}$. It is well known that OFDM technique is very sensitive to the frequency offset between the transmitter and the receiver oscillators. In fact, in the presence of a frequency error , the sub-carriers loose their orthogonality and the system performance becomes degraded. Note that in the absence of additive noise, the frequency offset must satisfy $|\epsilon| < 1.3.10^{-2}$ in order to obtain an effective signal-to-noise ratio of 30 dB or higher [6]. The main objective of this paper is to derive a frequency estimation algorithm for the OFDM signals with M-PSK modulation scheme based on the maximum likelihood (ML) principle. Its performance is computed and compared with simulation results. It is observed that the resulting algorithm generates very accurate estimate even when the offset is high. At the end, it is also shown that the frequency estimator may be used in a tracking mode.

3 Maximum Likelihood (ML) Estimator for Frequency Offset

Consider an observation vector, containing $N+L$ samples within the ith OFDM frame,

$$r = [r(1), r(2), \ldots, r(N+L)]^T,$$

where $r(k)$ is as defined by (1). As previously stated, of the $N+L$ samples, the first L samples are the cyclic prefix or the exact copy of the last L samples. As a result, the following relation exists between the OFDM symbols $s(k)$, transmitted:

$$s(k) = s(k+N) \quad \text{for } k = 1, 2 \ldots L$$

where,

$$s(k) = s_i^{k-1} \exp[j(2\pi\epsilon k/N + \phi)] \quad \text{for } k = 1, 2 \ldots N \tag{2}$$

and s_i^{k-1}'s are the IFFT of the data sequence given by

$$s_i^{k-1} = (1/N) \sum_{n=0}^{N-1} a_i^n \exp(j2\pi kn/N)$$

Here, a_i^n denotes M-PSK symbols transmitted on the nth carrier during the ith OFDM frame, taking values in the set $\{e^{j\frac{2\pi r}{M}}, r = 0, 1, \ldots M-1\}$. The likelihood function for ϵ and the data sequence $\{a_i^n\}$, $L(\epsilon, \{a_i^n\})$ is the probability density function of the $N+L$ observed samples given the carrier frequency offset ϵ and $\{a_i^n\}$. Note that the liklihood function is independent of the phase rotation ϕ and therefore, it is not shown in the formulation for the sake of simplicity.

Thus the liklihood function can be written as,

$$L(\epsilon, \{a_i^n\}) = \prod_{k=1}^{L} p_n[r(k) - s(k)] \prod_{k=L+1}^{N} p_n[r(k) - s(k)] \prod_{k=N+1}^{N+L} p_n[r(k) - s(k)] \tag{3}$$

$$= \prod_{k=1}^{L} p_n[r(k) - s(k)] p_n[r(k+N) - s(k+N)] \prod_{k=L+1}^{N} p_n[r(k) - s(k)] \tag{4}$$

Here $p_n(.)$ denotes the complex Gaussian density function with variance σ_n^2. It can be shown that the last product term in (4) is independent of ϵ for at SNR and/or for large M values. Therefore, it may be dropped.Minimizing (4) with respect to ϵ is equivalent to maximizing

$$L(\epsilon, \{a_i^n\}) = \prod_{k=1}^{L} p_n[r(k) - s(k)]p_n[r(k+N) - s(k+N)]. \tag{5}$$

after averaging it over the data symbols. Then, using (1) and (2), Eq.(5) can be expressed as,

$$L(\epsilon, \{a_i^n\}) = \exp\left\{ \frac{2}{\sigma_n^2} \sum_{k=1}^{L} Re\left\{ [r(k)\exp(-j2\pi\epsilon k/N) + r(k+N)\exp(-j2\pi\epsilon(k+N)/N)]\, s_i^{*k-1} \right\} \right\} \tag{6}$$

where (*) denotes a complex conjugate. Substituting (2) for s_i^{*k-1} and defining

$$R(k) \equiv [r(k) + r(k+N)\exp(-j2\pi\epsilon)]\exp[-j2\pi\epsilon k/N - j2\pi(k-1)n/N],$$

the liklihood ratio can be simplified to

$$L(\epsilon, \{a_i^n\}) = \exp\left\{ \frac{2}{N\sigma_n^2} \left(\sum_{k=1}^{L} \sum_{n=0}^{N-1} Re[a_i^{*n} R(k)] \right) \right\} \tag{7}$$

The data dependence in (7) is removed by averaging the likelihood ratio over all possible values of $\{a_i^n\}$. For independent, identically distributed M-PSK data sequence, the averaged liklihood function takes the form

$$L(\epsilon) = E[L(\epsilon, \{a_i^n\})] = \prod_{n=0}^{N-1} \prod_{k=1}^{L} \frac{1}{M} \sum_{r=0}^{M-1} \exp\left[\frac{2}{N\sigma_n^2} Re\left\{ -j\frac{2\pi r}{M} R(k) \right\} \right] \tag{8}$$

By making the approximation,

$$e^x \approx 1 + x + \frac{x^2}{2}$$

which applies in (8) at low SNR, it can be shown that for a trial value of ϵ, say $\hat{\epsilon}$ the final log-liklihood ratio takes the form:

$$\Lambda(\hat{\epsilon}) = N \sum_{k=1}^{L} \ln\left\{ 1 + \frac{2}{N^2\sigma_n^4} [|r(k)|^2 + |r(k+N)|^2 + 2Re[r(k)r^*(k+N)e^{j2\pi\hat{\epsilon}}] \right\} \tag{9}$$

For low signal-to-noise ratio, the above expression reduces to,

$$\Lambda(\hat{\epsilon}) = N \sum_{k=1}^{L} \frac{2}{N^2\sigma_n^4} \left[|r(k)|^2 + |r(k+N)|^2 + 2Re[r(k)r^*(k+N)e^{j2\pi\hat{\epsilon}} \right] \tag{10}$$

if $\nu = \sum_{k=1}^{L} r(k)r^*(k+N)e^{j2\pi\epsilon}$ and $\rho = |r(k)|^2 + |r(k+N)|^2$ then,

$$\Lambda(\hat{\epsilon}) = C_1|\nu|\cos(2\pi\hat{\epsilon} + \angle\nu) + C_2\rho. \tag{11}$$

where C_1 and C_2 are constants, independent of $\hat{\epsilon}$. The ML estimate of $\hat{\epsilon}$ can be easily obtained from (11) by taking derivative with respect to ϵ and setting it to zero, as,

$$\hat{\epsilon}_{ML} = -\frac{1}{2\pi}\angle\left(\sum_{k=1}^{L} r(k)r^*(k+N)\right) \tag{12}$$

4 Estimation of Carrier Offset in Tracking Mode

Taking derivative of (9) with respect to $\hat{\epsilon}$ gives,

$$\frac{d\Lambda(\hat{\epsilon})}{d\hat{\epsilon}} = N\sum_{k=1}^{L}\frac{(-8\pi/N^2\sigma_n^4)Im\{q_k\}}{1+2/N^2\sigma_n^4\,(\Phi_k+Re\{q_k\})} \tag{13}$$

where, $q_k = r(k)r^*(k+N)\exp(j2\pi\hat{\epsilon})$ and $\Phi_k = |r(k)|^2 + |r(k+N)|^2$. From (13), by taking $L=1$, we can derive an iterative algorithm for the optimization of $\hat{\epsilon}$,

$$\hat{\epsilon}_{k+1} = \hat{\epsilon}_k + \gamma E_k \tag{14}$$

where $0<\gamma<1$ is a positive constant, $\hat{\epsilon}_{k+1}$ and $\hat{\epsilon}_k$ are the estimation of $\hat{\epsilon}$ at the kth and $(k+1)$th iteration, respectively. E_k represents the error control signal given by

$$E_k = \frac{-Im\{q_k\}}{1+(2/N^2\sigma_n^4)(\Phi_k+Re\{q_k\})} \tag{15}$$

The above explained algorithm can be implemented using an automatic frequency control loop, as shown in Fig. 2. For the tracking mode also an OFDM symbol of length $N+L$ is needed at a time. The offset is tracked with the first and last L samples with taking only a pair of $r(k)$ and $r(k+N)$ during each track to obtain an $\hat{\epsilon}_k$. The rest of the samples are corrected with the most recent offset $\hat{\epsilon}_k$ which is found when $k=L$.

5 Mean Squared Performance of the Frequency Offset Estimator

The mean value of the derivative of the log-liklihood ratio is helpful to characterize the performance of the estimator. The following quantities are helpful in this respect.

$$W(\hat{\epsilon}) \cong \frac{\delta\Lambda}{\delta\hat{\epsilon}} \tag{16}$$

$$F(\hat{\epsilon}) \cong E[W(\hat{\epsilon})] \tag{17}$$

$$A \cong \frac{\delta F}{\delta\hat{\epsilon}}|_{\hat{\epsilon}=\epsilon} \tag{18}$$

The function $F(.)$ is often referred to as the "S-curve" or "discrimination characteristic" of the estimator. Let, $\hat{\epsilon}_{ML}$ be the maximum likelihood estimation, then for the case of relatively small jitter the approximation of variance is,

$$Var[\hat{\epsilon}_{ML}-\epsilon] = A^{-2}Var[W(\epsilon)] = \frac{E[W^2(\hat{\epsilon})]}{A^2}|_{\hat{\epsilon}=\epsilon} \tag{19}$$

From Eq.(11) it follows that,

$$W(\hat{\epsilon}) = \frac{\delta\Lambda}{\delta\hat{\epsilon}} = -2\pi|\nu|\sin(2\pi\hat{\epsilon}+\angle\nu), \tag{20}$$

from which, it is evident that,

$$E[W(\hat{\epsilon})] = -2\pi Im\left\{E[\nu]e^{j2\pi\hat{\epsilon}}\right\} \tag{21}$$

and,

$$E[W^2(\hat{\epsilon})] = -2\pi\left(E[|\nu|^2] - Re\left\{E[\nu^2]e^{j4\pi\hat{\epsilon}}\right\}\right) \tag{22}$$

The expression for A is derived to be,

$$A = -4\pi^2\sigma_s^2 L \tag{23}$$

where, $\sigma_s^2 = E[|s(k)|^2]$. Expressions for $E[|\nu|^2]$ and $E[\nu^2]$ are derived in Appendix-A and when they are substituted in (22) the following expression is obtained,

$$E[W^2(\hat{\epsilon})] = 2\pi^2 L[\sigma_n^4 + 2\sigma_s^2\sigma_n^2]. \tag{24}$$

Substituting the expressions for $E[W^2(\hat{\epsilon})]$ and A in (19) yields,

$$Var[\hat{\epsilon}_{ML} - \epsilon] = \frac{1 + 2SNR}{8\pi^2 L(SNR)^2} \tag{25}$$

where, $SNR = \sigma_s^2/\sigma_n^2$

6 Simulations

In order to simulate a baseband OFDM system, the following parameters were selected for frequency offset estimation: Total number of sub-channels, $N = 256$. Number of levels in M-ary PSK signaling, $M = 16$. Normalized carrier frequency offset, $\epsilon = 0.25$. Length of cyclic prefix, $L = 4$, $L = 8$ and $L = 15$ and only additive white Gaussian noise is present. Eq. (25) gives the analytical expression for the variance of the carrier offset estimation as a function of cyclic prefix L and SNR. Fig. 1 shows the comparison of analytical curves with the simulated curves. These curves indicate an excellent agreement between the analytical and simulation results. Note that the simulation performs better for longer prefix length because of greater averaging.

6.1 Performance of Estimator in Tracking Mode

To simulate the estimation algorithm in tracking mode the block diagram in Fig. 2 is used. In tracking mode the estimated error offset of one symbol is fed to the next symbol as to have better convergence to the original value of 0.25. The signaling is always chosen to be 16-PSK. It is observed that for values of $\gamma > 10^{-5}$ the estimated value for ϵ fluctuates very much and even gives negative values. So, the system is simulated for γ values of $10^{-5}, 10^{-6}$ and 10^{-7}. ¿From Fig. 3 it is evident that for the first 500 symbols, when $\gamma = 10^{-7}$, the system does not reach steady state condition and it would take too much time to track and converge to actual value. With $\gamma = 10^{-5}$ error offset is tracked to be 0.25 only after about 20 symbols but there are too much fluctuations after that. However, for $\gamma = 10^{-6}$ it takes about 200 symbols to reach the steady state condition but once it has been reached there is no more jitters. So, the optimal value of γ is chosen to be 10^{-6} for this particular setup. Since in tracking mode, the error is tracked from sample to sample in each symbol and then the last estimated value of offset is feeded as the initial estimation for the next symbol, it is expected that performance of estimation would be better than the forward estimation mode. Also, as the estimated value for offset in the transient region (first 200 symbols) is discarded, variance for different signal to noise ratio is much lower that normal mode case. Simulation is done for 10,000 symbols in each case and $L = 4$ is chosen for comparison. Fig. 4 shows that variance in normal mode is higher than variance in tracking mode.

6.2 Performance of the OFDM System

Performance of the OFDM system is determined in terms of the symbol error rate (SER) plot for different signal to noise ratios. To correct frequency error offset, a closed loop realization of the algorithm is carried out. The SER of the OFDM system is calculated at different SNR and lengths of cyclic prefix. As variance of estimation becomes better with higher SNR and longer prefix, SER is also low under these conditions as expected. Fig. 5 shows the comparison of SER versus SNR as the cyclic prefix change. The above Fig. 5 also compares these curves with the ideal system performance for which there is no frequency offset. It can be seen that for $L = 15$, as signal power becomes quite low the system is a bit unstable for signal to noise ratio greater than 14 dB. For the tracking mode, since it takes first 200 symbols track to correct offset, these 200 symbols are discarded. So, in tracking mode the system behaves almost as if there was no frequency offset as after discarding the first 200 symbols the estimated error is very near the actual error. Fig. 6 shows the comparison of forward estimation mode and the tracking mode performance of the OFDM system along with ideal system. Tracking mode curve almost coincides with the ideal system curve at the cost of discarding the first 200 symbols.

7 References

[1] W. Y. Zou and Y. Wu, " COFDM: an overview",*IEEE Trans. Broadcasting*, vol. 41, pp. 1-8, March 1995.

[2] J. A. C. Bingham," Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come," *IEEE Commun. Mag.*, V.28, pp. 5-14, May1990.

[3] L. Wei and C. Schlegel, " Synchronization requirements for multi-user OFDM on satellite mobile and two-path Rayleigh fading channels",*IEEE Trans. Commun.*, vol. 43, pp. 887-895, Feb./Mar./Apr. 1995.

[4] T. Pollet, M. van Bladel and M. Moeneclaey, " BER sensivity of OFDM systems to carrier frequency offset and Wiener process noise", *IEEE Trans. Commun.*, vol. 43, pp. 191-193, Feb./Mar./Apr. 1995.

[5] W. D. Warner and C. Leung, " OFDM/FM frame synchronization for mobile radio data communication", *IEEE Trans. Veh. Technology*, vol. 42, pp. 302-313, August 1993.

[6] P. J. Tourtier, R. Monnier and P. Lopez, " Multicarrier modem for digital HDTV terrestial broadcasting", *Signal Processing: Image Commun.*, vol. 5, n0. 5-6, pp.379-403, december 1993.

[7] F. Daffara and A. Chouly, "Maximum Likelihood Frequency Detectors for Orthogonal Multicarrier Systems," *Proc. IEEE Int. Conf. Commun.*, pp. 766-771, May 1993.

[8] J. Beek, M. Sandell and P. O. Borjesson,"ML Estimation of Time and Frequency Offset in OFDM Systems," *IEEE Transactions on Signal Processing*, Vol. 45, NO. 7, July 1997.

[9] P. H. Moose, " A technique for orthogonal frequency division multiplexing frequency offset correction", *IEEE Trans. Commun.*, vol. 42, pp. 2908-2914, Oct. 1994.

[10] F. Daffara and O. Adami, " A new frequency detector for orthogonal multicarrier transmission techniques", in *Proc. Vehic. Technol. Conf.*, Chicago, IL, vol.2 July2, 1995, pp. 804-809.

[11] R. W. Chang," Orthogonal Frequency Division Multiplexing," *U. S. Patent 3,488,445*, filed 1966, issued Jan. 6, 1970

Appendix-A
Derivations of $E[|\nu|^2]$ and $E[\nu^2]$

¿From (??) it can be stated that,

$$E[|\nu|^2] = \sum_k \sum_l E[r(k)r^*(k+N)r^*(l)r(l+N)]$$

where

$$E[r(k)r^*(k+N)r^*(l)r(l+N)] = \begin{cases} E[r(k)r^*(k+N)]E[r^*(l)r(l+N)] & \text{for } k \neq l \\ E[|r(k)|^2|r(k+N)|^2]] & \text{for } k = l \end{cases}$$

For the first case,

$$E[r(k)r^*(k+N)]E[r^*(l)r(l+N)] = \sigma_s^2 e^{-j2\pi\epsilon}\sigma_s^2 e^{j2\pi\epsilon} = \sigma_s^4$$

For the second case,

$$E[|r(k)|^2|r(k+N)|^2] = E[|s(k)|^4] + \sigma_n^4 + 2\sigma_s^2\sigma_n^2$$

where, $\sigma_n^2 = E[|n(k)|^2]$ is the noise variance. So,

$$E[|\nu|^2] = L(L-1)\sigma_s^4 + L\left(E[|s(k)|^4] + \sigma_n^4 + 2\sigma_s^2\sigma_n^2\right)$$

Again,from (26)

$$E[\nu^2] = \sum_k \sum_l E[r(k)r^*(k+N)r(l)r^*(l+N)]$$

where

$$E[r(k)r^*(k+N)r(l)r^*(l+N)] = \begin{cases} E[r(k)r^*(k+N)]E[r(l)r^*(l+N)] & \text{for } k \neq l \\ E[r(k)^2 r^*(k+N)^2]] & \text{for } k = l \end{cases}$$

Similarly, for the first case,

$$E[r(k)r^*(k+N)]E[r(l)r^*(l+N)] = \sigma_s^4 e^{-j4\pi\epsilon}$$

For the second case,

$$E[r(k)^2 r^*(k+N)^2]] = E[|s(k)|^4]e^{-j4\pi\epsilon}$$

So,

$$E[\nu^2] = L(L-1)\sigma_s^4 e^{-j4\pi\epsilon} + L[E[|s(k)|^4]e^{-j4\pi\epsilon}.$$

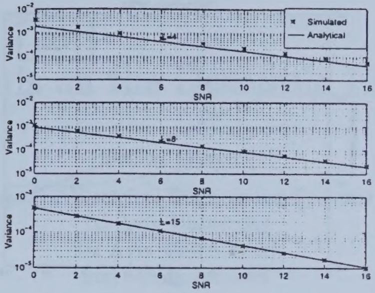


Figure 1: Comparison of analytical and simulated curves for variance of estimator

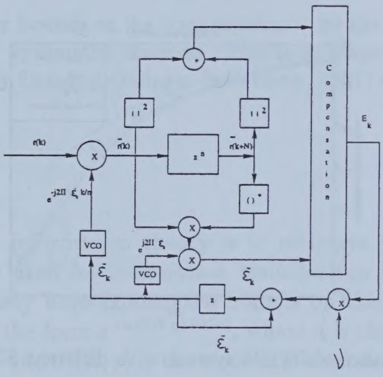

r(k)
X
r̄(k)
z^n
r̄(k+N)
| |²
| |²
+
( )*
X
X
e^{-j2Π ξ k/n}
VCO
VCO
e^{j2Π ξ}
ξ̄_k
ξ̄_k
z
+
X
ξ̄_k
E_k
Compensation

Figure 2: Tracking of Frequency Offset

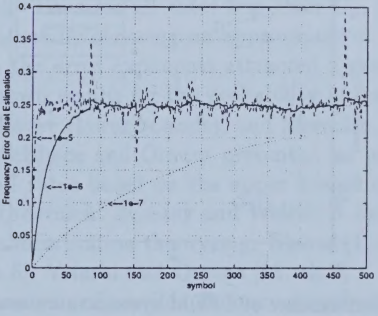


Figure 3: Tracking of error offset when $\gamma = 10^{-5}, 10^{-6}, 10^{-7}$, $SNR = 10dB$

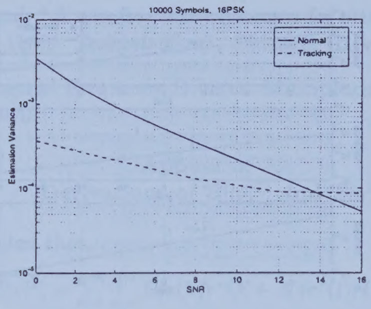


Figure 4: Comparison of estimator variance in normal and tracking mode

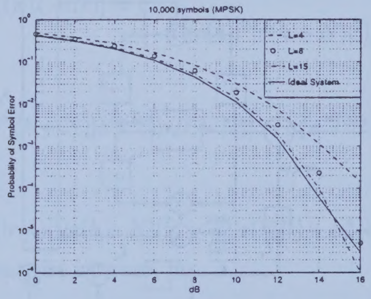


Figure 5: Performance of OFDM system with different SNR and cyclic prefix

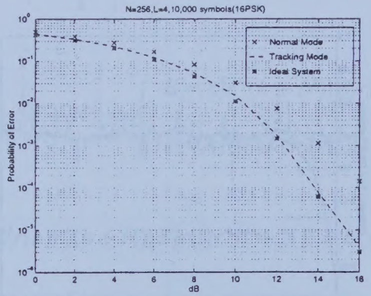


Figure 6: Comparison of performance of OFDM system in normal and tracking mode, $L = 4$

New Lower Bounds on the Decoding Error Probability

Simon Litsyn *
Dept. of Electrical Engineering-Systems
Tel Aviv University
Tel Aviv 69978 Israel
litsyn@eng.tau.ac.il

Abstract

We derive new upper bounds on the error exponents for the maximum likelihood decoding in the binary symmetric channels. This is an improvement on the best earlier known bounds by Shannon-Gallager-Berlekamp (1967) and McEliece-Omura (1977).

1 Introduction

A classical problem of the information theory is to estimate probabilities of decoding error when a block code is used for information transmission over a binary symmetric channel (BSC). We will study here exponential bounds on the performance of the best codes. These bounds are of the form $e^{-nE(R,p)+o(n)}$, where n is the length of a code, $o(n)/n$ tends to zero when n grows, and $E(R,p)$ is the error exponent depending only on the code rate R and the transition probability p. The function $E(R,p)$ was thoroughly studied since 1948, when Shannon proved in his seminal paper [15] that the error exponent for the maximum likelihood decoding is positive in the interval $0 \le R < C_p$, where C_p is the capacity of the channel. Thus for all rates less than C_p, the error probability can be made as small as one wishes by choosing an appropriate n. The problem of deriving lower and upper bounds on the error exponents attracted a great deal of attention since then. The best currently known results for the probability of decoding error were derived in the middle sixties by Gallager (lower bounds), and Shannon, Gallager and Berlekamp (upper bounds). In 1977 McEliece and Omura presented an improvement of the upper bounds in the range of small rates based on the upper bound on the minimum distance of codes due to McEliece, Rodemich, Rumsey and Welch. A description of these bounds appear in every textbook on information theory, e.g. Blahut [1, ch.10], Csisár and Körner [2, ch.2, §5], Gallager [6, ch.5], Viterbi and Omura [18, ch.3].

In this paper, using an analysis of possible distance distributions of codes, we improve the upper bounds on the error exponents, or, which is the same, the lower bounds on the maximum likelihood decoding error probability.

To state the problem rigorously we need some notations. Let $\mathbf{F}^n = \{0,1\}^n$, be the space of binary n-tuples endowed with the Hamming metric,

$$d(\mathbf{x},\mathbf{y}) = \left|\{i : x_i \neq y_i\}\right|, \tag{1}$$

---

*Partly supported by DIMACS, Center for Discrete Mathematics and Theoretical Computer Science, Rutgers University where $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{y} = (y_1, \ldots, y_n), \mathbf{x} \in \mathbf{F}^n, \mathbf{y} \in \mathbf{F}^n$. The Hamming weight of an $\mathbf{x} \in \mathbf{F}^n$ is

$$wt(\mathbf{x}) = |\{i : x_i \neq 0\}|.$$

Let $C = C(n, |C|, d) \subseteq \mathbf{F}^n$ be a code of size $|C|$ and minimum distance

$$d = d(C) = \min_{\mathbf{c}, \tilde{\mathbf{c}} \in C, \mathbf{c} \neq \tilde{\mathbf{c}}} d(\mathbf{c}, \tilde{\mathbf{c}}).$$

Let $d(K, n)$ be the maximal possible minimum distance of a code of length n and size K. The rate $R(C)$ and relative distance $\delta(C)$ of a code $C(n, |C|, d)$ are defined as (all logarithms are base 2):

$$\begin{aligned} R = R(C) &= \log |C|/n, \\ \delta = \delta(C) &= d(C)/n. \end{aligned}$$

Let $H(x)$ be the binary entropy function

$$H(x) = -x \log x - (1-x) \log(1-x), \ x \in [0,1],$$

and $H^{-1}(y), y \in [0, 1/2]$, stand for its inverse function. Define for $0 \leq R \leq 1$, $K = \lceil 2^{Rn} \rceil$,

$$\delta(R) = \limsup_{n \to \infty} \frac{d(n, K)}{n}.$$

Then, by the Gilbert-Varshamov bound [8, 17],

$$\delta(R) \geq \delta_{GV}(R) := H^{-1}(1 - R). \tag{2}$$

On the other hand, the McEliece-Rodemich-Rumsey-Welch (linear programming) bound [14] (see [10] for the form of the bound below) yields that

$$\delta(R) \leq \delta_{LP}(R) := \min_{0 \leq \beta \leq \alpha \leq 1/2, \ H(\alpha) - H(\beta) = 1 - R} 2 \frac{\alpha(1-\alpha) - \beta(1-\beta)}{1 + 2\sqrt{\beta(1-\beta)}}. \tag{3}$$

Let $R_0 = H(\beta_0)$, where $\beta_0 \approx 0.0545...$ is the only root in $(0, 1/2)$ of the equation

$$(1 - 2x)\left(2 + \frac{1}{\sqrt{x(1-x)}}\right) - 2 \ln\left(\frac{1-x}{x}\right) = 0.$$

Then for $R \leq R_0 \approx 0.3052...$, the minimum in (3) is attained when $\alpha = 1/2$, and (3) gets especially simple form:

$$\delta(R) \leq \delta_{LP1}(R) := \frac{1}{2} - \sqrt{H^{-1}(R)(1 - H^{-1}(R))}. \tag{4}$$

In the axis of relative distance, the two bounds coincide for

$$\delta \geq \delta_{LP}(R_0) \approx 0.272... \tag{5}$$

otherwise, the bound (3) is tighter.

The bound (4) is also valid for all $R \in [0, 1]$. A long-standing conjecture is that the Gilbert-Varshamov bound (2) is actually tight.

A BSC is described by the probability p that the input symbol (0 or 1) and the corresponding output symbol (0 or 1) differ. We assume that $0 \le p \le 1/2$. Given a code $\mathcal{C} \subseteq \mathbf{F}^n$ we can partition $\mathbf{F}^n$ to decode regions $U(\mathbf{c})$, $\mathbf{c} \in \mathcal{C}$. A vector $\mathbf{z}$ belongs to $U(\mathbf{c})$ if there is no other code word $\tilde{\mathbf{c}} \ne \mathbf{c}$ such that $d(\mathbf{z}, \tilde{\mathbf{c}}) \le d(\mathbf{z}, \mathbf{c})$. If $\mathbf{z}$ is such that there is no codeword $\tilde{\mathbf{c}} \ne \mathbf{c}$ such that $d(\mathbf{z}, \tilde{\mathbf{c}}) < d(\mathbf{z}, \mathbf{c})$, but there are several, say ℓ, $\ell \ge 1$, code words, $\tilde{\mathbf{c}}^{(1)}, \ldots, \tilde{\mathbf{c}}^{(\ell)}$, satisfying

$$d(\mathbf{z}, \tilde{\mathbf{c}}) = d(\mathbf{z}, \tilde{\mathbf{c}}^{(1)}) = \ldots = d(\mathbf{z}, \tilde{\mathbf{c}}^{(\ell)}),$$

then we assign $\mathbf{z}$ arbitrarily to anyone from $U(\mathbf{c}), U(\tilde{\mathbf{c}}^{(1)}), \ldots, U(\tilde{\mathbf{c}}^{(\ell)})$.

Then the average probability of decoding error in BSC (provided that all code words are equiprobable) is

$$P_{de}(\mathcal{C}, p) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \sum_{z \in F^n \setminus U(c)} p^{d(c,z)} (1-p)^{n-d(c,z)}. \tag{6}$$

The distance distribution of a code $\mathcal{C}$ is an $(n+1)$-tuple $\mathcal{B}(\mathcal{C}) = \mathcal{B} = (\mathcal{B}_0, \ldots, \mathcal{B}_n)$, defined by

$$\mathcal{B}_i = \frac{1}{|\mathcal{C}|} \Big| \{\mathbf{c}, \tilde{\mathbf{c}} : \mathbf{c}, \tilde{\mathbf{c}} \in \mathcal{C}, d(\mathbf{c}, \tilde{\mathbf{c}}) = i\} \Big|.$$

Assume $K = \lceil 2^{Rn} \rceil$ if $R > 0$, and $K = n + 1$ if $R = 0$ (the last condition guarantees continuity at $R = 0$). Let

$$E_{de}(R, p) = \limsup_{n \to \infty} \; -\frac{1}{n} \log \min_{\mathcal{C} \subseteq F^n, |\mathcal{C}| = K} P_{de}(\mathcal{C}, p), \tag{7}$$

For positive f and g, we use notation $f \succeq g$ if $f \ge g(1 + o(1))$ and $o(1)$ tends to 0 when n grows. Analogously, we write $\preceq$ and $\simeq$ for $f \le g(1 + o(1))$ and $f = g(1 + o(1))$.

Define

$$T(x, y) = -x \log y - (1 - x) \log(1 - y).$$

Notice, that $H(x) = T(x, x)$.

About the probability of decoding errors the following is known. The best lower bounds are due to Elias, 1956, and Gallager, 1963, 1965, 1968,

$$E_{de}(R, p) \succeq \Bigg\{$$

$$\begin{array}{rll} T(\delta_{GV}(R), p) + R - 1 & \text{if } R_{crit} \le R \le C_p & [3] \qquad (8) \\ 1 - \log\left(1 + \sqrt{4p(1-p)}\right) - R & \text{if } R_{min} \le R \le R_{crit} & [5, 6] \qquad (9) \\ -\delta_{GV}(R) \log \sqrt{4p(1-p)} & \text{if } 0 \le R \le R_{min} & [5, 6] \qquad (10) \end{array}$$

Here

$$R_{min} = 1 - H\left(\frac{\sqrt{4p(1-p)}}{1 + \sqrt{4p(1-p)}} \right)$$

$$R_{crit} = 1 - H\left(\frac{\sqrt{p}}{\sqrt{p} + \sqrt{1-p}} \right)$$

We address the bound (8) as the *lower sphere packing bound*, the bound (9) as the *lower straight-line bound*, and the bound (10) as the *lower minimum distance bound*.

The upper bounds are due to Elias, 1956, Shannon, Gallager and Berlekamp, 1967, and McEliece and Omura, 1977 (based on the McEliece, Rodemich, Rumsey and Welch, 1977).

$$E_{de}(R,p) \preceq \{$$

$$T(\delta_{GV}(R),p)+R-1 \qquad \text{if } 0 \le R \le C_p \qquad [3] \qquad (11)$$

$$-\delta_{LP}(R)\log\sqrt{4p(1-p)} \qquad \text{if } 0 \le R \le C_p \qquad [16,13] \qquad (12)$$

$$T(\delta_{LP}(R_1),p)\frac{R_2-R}{R_2-R_1}+$$

$$+(T(\delta_{GV}(R_2),p)+R_2-1)\frac{R-R_1}{R_2-R_1} \quad \text{for any} R_1 \le R \le R_2 \quad [16,13] \qquad (13)$$

Bound (11) proves that (8) is tight in the range $[R_{crit}, C_p]$. Bound (13) is an application of the following result due to Shannon, Gallager and Berlekamp.

Lemma 1 *Any straight line between a point on any upper bound and a point on (11) is also an upper bound.*

Here it is applied to (12). This argument yields tightness of (8), (9) and (10), modulo the conjecture about the tightness of the Gilbert-Varshamov bound. We address the bounds (11), (12) and (13) as the *upper sphere-packing*, *upper minimum distance* and *upper straight-line* bounds, correspondingly.

There is still a gap between the lower and upper bounds, and it is a long-standing open problem how to close or, at least, decrease it. Since any improvement in the upper bounds on $\delta(R)$ leads to an improvement on the upper bounds for the exponents, the main efforts were applied in attempts to improve on (3). We quote Viterbi and Omura [18, p.184]:

> "The most likely improvement ... will come about as a result of an improvement in the upper bound on minimum distance."

However, we present here an alternative approach, which does not require improving upper bounds on the minimum distance, nevertheless leading to better upper bounds on the error exponents. The main idea is as follows. We prove that there necessarily should be a large enough component in the distance distribution. Further we demonstrate that the impact of this component results in an exponential factor, thus improving on the earlier known bounds. The new bounds on the exponents are summarized in the following theorem.

Theorem 1

$$E_{de}(R,p) \preceq \max_{\alpha,\beta,\xi,\delta} E_{\alpha,\beta,\xi,\delta},$$

where

$$E_{\alpha,\beta,\xi,\delta} = \min\Big(-\delta\log\sqrt{4p(1-p)}, -\tilde{\nu}-\xi\log\sqrt{4p(1-p)}\Big),$$

α,β, ξ and δ are such that

$$0 \le \beta \le \alpha \le 1/2, \quad H(\alpha)-H(\beta) = 1-R,$$

$$\xi \in \left[0, 2\frac{\alpha(1-\alpha)+\beta(1-\beta)}{1+2\sqrt{\beta(1-\beta)}}\right], \qquad \delta \in [0,\delta_{LP}(R)],$$

$$\tilde{\nu} = \min\Big(\nu, \xi + (1-\xi)H(p) - \max_{\eta\in[\delta p/2, \min(\delta/4, p(1-\xi))]} \Big(\delta H\left(\frac{2\eta}{\delta}\right)$$
$$+(\xi - \delta/2)H\left(\frac{\xi - 2\eta}{2\xi - \delta}\right) + (1-\xi-\delta/2)H\left(\frac{p(1-\xi)-\eta}{1-\xi-\delta/2}\right)\Big)\Big),$$

$$\nu = R - 1 + H(\beta) + 2H(\alpha) - 2q(\alpha,\beta,\xi/2) - \xi - (1-\xi)H\left(\frac{\alpha - \xi/2}{1-\xi}\right),$$

$$q(\alpha,\beta,\xi) = H(\beta) + \int_0^{\xi} \log\Bigg(\frac{(\alpha(1-\alpha) - y(1-2y) - \beta(1-\beta))}{2(\alpha-y)(1-\alpha-y)}$$
$$+\frac{\sqrt{(\alpha(1-\alpha)-y(1-2y)-\beta(1-\beta))^2 - 4(\alpha-y)(1-\alpha-y)y^2}}{2(\alpha-y)(1-\alpha-y)}\Bigg)\,dy.$$

2 A bound on the distance distribution

We present here the key result about distance distributions of codes yielding the improvement claimed in Theorem 1.

Let C be a code of length n with the distance distribution $\mathcal{B} = (\mathcal{B}_0 = 1, \mathcal{B}_1, \ldots, \mathcal{B}_n)$.

Theorem 2 *In every code C of big enough length n there exists a component of the distance distribution such that*

$$\frac{1}{n}\log \mathcal{B}_{2\xi n} \succeq \hat{R}_{2\xi}(C) \tag{14}$$

$$:= R(C) - 1 + H(\beta) + 2H(\alpha) - 2q(\alpha,\beta,\xi) - 2\xi - (1-2\xi)H\left(\frac{\alpha-\xi}{1-2\xi}\right)$$

for $\xi n > 0$ and

$$2\xi \in \left[0, 2\frac{\alpha(1-\alpha)+\beta(1-\beta)}{1+2\sqrt{\beta(1-\beta)}}\right], \tag{15}$$

where $q(\alpha,\beta,\xi)$ is defined by

$$q(\alpha,\beta,\xi) = H(\beta) + \int_0^{\xi} \log\Bigg(\frac{(\alpha(1-\alpha) - y(1-2y) - \beta(1-\beta))}{2(\alpha-y)(1-\alpha-y)} \tag{16}$$
$$+\frac{\sqrt{(\alpha(1-\alpha)-y(1-2y)-\beta(1-\beta))^2 - 4(\alpha-y)(1-\alpha-y)y^2}}{2(\alpha-y)(1-\alpha-y)}\Bigg)\,dy,$$

α and β are such that $0 \le \beta \le \alpha \le 1/2$, $H(\alpha) - H(\beta) \ge 1 - R(C)$.

Notice that the upper limit of the interval in (15) is at least $\delta_{LP}(R(C))$ (compare the expresssion to (3)).

The last theorem along with some quite involved combinatorial arguments yield the new upper bound on the error exponent. We omit details of the proof.

The new bound improves on the earlier known bounds in all the interval of rates where the exact values of the exponent are not known, i.e. in the interval $[0, R_{crit}]$.

References

[1] R. E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, 1987.

[2] I. Csisziár and J. Körner, *Information Theory. Coding Theorems for Discrete Memoryless Systems*, Akademiai Kiado, 1981.

[3] P. Elias, "Coding for two noisy channels," in: *Information Theory*, Academic Press, 1956, pp.61–74.

[4] R. G. Gallager, *Low-density Parity-check Codes*, MIT Press, 1963.

[5] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol.11, 1965, pp.3–18.

[6] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley and Sons, 1968.

[7] R. G. Gallager, "The random coding bound is tight for the average code," *IEEE Trans. Inform. Theory*, vol.19, 1973, pp.244–246.

[8] E. N. Gilbert, "A comparison of signalling alphabets," *Bell Syst. Tech. J.*, vol.31, no.3, 1952, pp.504–522.

[9] R. W. Hamming, "Error detecting and error correcting codes," *Bell Syst. Tech. J.*, vol.29, 1950, pp.147–160.

[10] V. I. Levenshtein, "Bounds for packings of metric spaces and some applications," *Probl. Kibernetiki*, vol.40, 1983, pp.43–110 (In Russian).

[11] J. H. van Lint, *Introduction to Coding Theory*, Spinger-Verlag, 1992.

[12] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, 1977.

[13] R. J. McEliece and J. K. Omura, "An improved upper bound on the block coding error exponent for binary-input discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol.23, no.5, 1977, pp.611–613.

[14] R. J. McEliece, E. R. Rodemich, H. Rumsey, Jr., and L. R. Welch, "New upper bounds on the rate of a code via the Delsarte-MacWilliams inequalities," *IEEE Trans. Inform. Theory*, vol.23, no.2, 1977, pp.157–166.

[15] C. E. Shannon, "A mathematical theory of communications," *Bell Syst. Tech. J.*, v.27, 1948, pp.379–423, 623–656.

[16] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels," *Information and Control*, vol.10, 1967, pp.65–103 (PartI), pp.522–552 (Part II).

[17] R. R. Varshamov, *An estimate for the number of signals in codes correcting errors*, *Dokl. AN SSSR*, vol.117, no.5, 1957, pp.739–741.

[18] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw-Hill, 1979.

A quantum decoding algorithm of the simplex code

Alexander Barg
Bell Laboratories, Lucent Technologies
600 Mountain Avenue, 2C-375
Murray Hill, NJ 07974
E-mail abarg@research.bell-labs.com

Shiyu Zhou
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104-6389
E-mail shiyu@central.cis.upenn.edu

Abstract

We suggest a quantum computation performing the decoding of the simplex code of length n with time complexity $O(\log^2 n)$. This is exponentially faster than the corresponding classical algorithm.

1 Introduction

Recently it has been shown [8], [6] that quantum algorithms can solve search problems significantly faster than classical computations. These discoveries generated a large amount of research (see [9]). The subject of this paper is a study of the decoding problem of binary linear codes.

Let $C \subseteq \mathbf{Z}_2^n$ be a binary linear code. The decoding mapping $\delta : \mathbf{Z}_2^n \to C$ is defined as follows:

$$\delta(\boldsymbol{x}) = \boldsymbol{c} \quad \Leftrightarrow \quad \mathsf{dist}\,(\boldsymbol{x}, \boldsymbol{c}) = \mathsf{dist}\,(\boldsymbol{x}, C).$$

If the last condition is satisfied for several codevectors $\boldsymbol{c}$, the value of δ is chosen arbitrarily from them. The algorithmic decoding problem is to implement the function δ. It is known [3] that for arbitrary (linear) code this problem is NP-hard. On the other hand, as an immediate consequence of Grover's quantum search algorithm [6], the decoding of any linear code of length n can be solved on a quantum computer in time $O(n|C|^{1/2})$, which is essentially optimal following a result in [2].

In this paper we take C to be the *binary simplex code.* The classical decoding algorithm has both time and space complexity $O(n \log n)$ [5]. It turns out that there exists a quantum probabilistic algorithm that provides a considerable (exponential) speedup of decoding. However, we have to somewhat restrict the mapping. Namely, let $\boldsymbol{c} \in C$, let $\theta \in (0, 1)$, and define

$$\begin{aligned} D_\theta(\boldsymbol{c}) &:= \{\boldsymbol{x} \in \mathbf{Z}_2^n : \forall_{\boldsymbol{c}' \in C, \boldsymbol{c}' \neq \boldsymbol{c}}\ \mathsf{dist}\,(\boldsymbol{c}, \boldsymbol{x}) - \mathsf{dist}\,(\boldsymbol{c}', \boldsymbol{x}) \leq -\theta n\}, \qquad (1) \\ E_\theta(\boldsymbol{c}) &:= \{\boldsymbol{x} \in \mathbf{Z}_2^n : \forall_{\boldsymbol{c}' \in C, \boldsymbol{c}' \neq \boldsymbol{c}}\ \mathsf{dist}\,(\boldsymbol{c}, \boldsymbol{x}) + \mathsf{dist}\,(\boldsymbol{c}', \boldsymbol{x}) \leq (1-\theta) n\}. \qquad (2) \end{aligned}$$

Let $D_\theta(C) = \cup_{\boldsymbol{c} \in C} D_\theta(\boldsymbol{c})$ and $E_\theta(C) = \cup_{\boldsymbol{c} \in C} E_\theta(\boldsymbol{c})$.

Let δ_θ be a restriction of δ to the intersection $D_\theta(C) \cap E_\theta(C)$. The algorithm implementing δ_θ that we construct, has time complexity $O(\log^2 n)$ and the probability of decoding error n^{-c}, where the exponent depends on θ.

2 Preliminaries

Let $n = 2^m$, where m is a positive integer. We shall use both numbers $v \in \{0, 1, \dots, n-1\}$ and their binary expansions. To tell one from the other, we denote by $\vec{v}$ the vector of coefficients of the binary expansion of v. To any vector $\boldsymbol{x} \in \mathbf{Z}_2^n$ we associate a function

$$\begin{array}{rccc} \boldsymbol{f}_x : & \mathbf{Z}_2^m & \rightarrow & \mathbf{Z}_2 \\ & \vec{v} & \mapsto & \boldsymbol{x}_v \end{array}$$

such that its value on $\vec{v}$ is equal to the vth bit of $\boldsymbol{x}$. This defines a bijection between the set of Boolean functions of m arguments and $\mathbf{Z}_2^n$. By abuse of notation, below we use bold symbols to denote both the function and the vector of its values. For instance, $\boldsymbol{c}(\vec{v})$ means the vth bit of the vector $\boldsymbol{c} \in \mathbf{Z}_2^n$.

The functional representation of $\mathbf{Z}_2^n$ is convenient in the study of Reed-Muller codes and their subcodes since one can define the code by specifying a subset of functions. In particular, the dual space $(\mathbf{Z}_2^m)^*$, i.e., the space of all linear functions $\{\sum_{i=1}^m u_i v_i\}$ on $\mathbf{Z}_2^m$, defines an $[n = 2^m, m, 2^{m-1}]$ linear code C (the simplex code extended by an all-zero column). In other words, a codevector $\boldsymbol{c} \in C$ has the form $(c_0, c_1, \dots, c_{n-1})$, where $c_v = \langle \vec{u}, \vec{v} \rangle$ and $\langle \cdot, \cdot \rangle$ is the standard dot product on $\mathbf{Z}_2^m$. The encoding procedure is given by the isomorphism $(\mathbf{Z}_2^m)^* \cong C \subset \mathbf{Z}_2^n$. The codevector corresponding to a message $\vec{u} \in (\mathbf{Z}_2^m)^*$ (the image of the covector $\vec{u}$ under this isomorphism) is denoted by $\boldsymbol{c}_{\vec{u}}$. By the Plotkin bound, the covering radius of the simplex code is $\rho(C) = 2^{m-1} - 1$.

An obvious consequence of the definition of C is as follows: for every 2 codevectors $\boldsymbol{c}_1 \neq \boldsymbol{c}_2$ we have

$$\mathsf{dist}\,(\boldsymbol{c}_1, \boldsymbol{c}_2) = 2^{m-1} = \frac{n}{2}. \tag{3}$$

Given any vector $\boldsymbol{r} \in \mathbf{Z}_2^n$, define the function $F_{\boldsymbol{r}} : \mathbf{Z}_2^m \rightarrow \{-1, 1\}$ by

$$F_{\boldsymbol{r}}(\vec{v}) = (-1)^{\boldsymbol{r}(\vec{v})}.$$

Definition 2.1 *The function*

$$\begin{array}{rccl} \widehat{F}_{\boldsymbol{r}} : & \mathbf{Z}_2^m & \rightarrow & \mathbf{C} \\ & \vec{u} & \mapsto & \sum_{\vec{v} \in \mathbf{Z}_2^m} F_{\boldsymbol{r}}(\vec{v})(-1)^{\langle \vec{u}, \vec{v} \rangle} \end{array}$$

is called the Walsh-Hadamard transform of the function $F_{\boldsymbol{r}}$ (equivalently, of the vector $\vec{u} \in \mathbf{Z}_2^m$).

Note that one can think of $\widehat{F}_{\boldsymbol{r}}$ as of the mapping defined on $(\mathbf{Z}_2^m)^*$.

We can also rewrite the definition of $\widehat{F}_{\boldsymbol{r}}$ as follows:

$$\widehat{F}_{\boldsymbol{r}}(\vec{u}) = \sum_{\vec{v} \in \mathbf{Z}_2^m} (-1)^{\boldsymbol{r}(\vec{v}) + \langle \vec{u}, \vec{v} \rangle}.$$

This implies the following proposition, which forms the basis for the decoding algorithm of C.

Proposition 2.1 *For any $\boldsymbol{r} \in \mathbf{Z}_2^n$ and any $\vec{u} \in (\mathbf{Z}_2^m)^*$,*

$$\widehat{F}_{\boldsymbol{r}}(\vec{u}) = n - 2\mathsf{dist}\,(\boldsymbol{r}, \boldsymbol{c}_{\vec{u}}).$$

To define a quantum-mechanical implementation of this algorithm, we shall represent vectors as states of quantum systems. An individual bit is represented by a 2-dimensional complex linear space, called qubit. The computation is built as a sequence of unitary operations composed from the two basic blocks, the rotation by 45°

$$W_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

and the conditional sign-shift transform

$$P_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Qubits are combined into larger quantum systems, which is expressed by the direct (tensor) product of the corresponding linear spaces. Likewise, operators acting on these systems will be formed by taking tensor products of the basic operations. For instance, the Walsh-Hadamard transform on a system of m qubits is given by $W_2^m := W_2^{\otimes m}$, etc.

3 The algorithm

The algorithm that we propose depends on the θ introduced in (1)-(2), and a parameter c_0 (a positive integer), which accounts for a tradeoff between the probability of decoding error P_e and the complexity. Namely, if the algorithm is iterated $t = c_0 \ln n$ times, then $P_e = n^{-c}$, where c depends both on c_0 and θ (see Lemma 4.2 below).

Let $\boldsymbol{r} \in \mathbf{Z}_2^n$ be the received vector. Our decoding algorithm will most probably output the closest codevector of C if $\boldsymbol{r} \in D_\theta \cap E_\theta$. In general, we do not have a more intuitive description of the decoding region. However, it is not difficult to see that

$$\Big(\bigcup_{\boldsymbol{c} \in C} S_{n(\frac{1}{4}-\theta)}(\boldsymbol{c}) \Big) \subset (D_\theta \cap E_\theta),$$

where $S_u(\boldsymbol{c})$ is the metric ball of radius u centered at $\boldsymbol{c}$, and that the inclusion is proper. Here are some consequences of definitions (1)-(2) that prove useful below. If $\boldsymbol{r} \in D_\theta \cap E_\theta$ and $\boldsymbol{c} = \delta_\theta(\boldsymbol{r})$ is the codevector closest to $\boldsymbol{r}$, then we can solve the inequalities with respect to $\mathsf{dist}\,(\boldsymbol{r}, \boldsymbol{c})$ to obtain

$$\mathsf{dist}\,(\boldsymbol{r}, \boldsymbol{c}) \le n(\frac{1}{2} - \theta). \tag{4}$$

Next, by (3) we obtain, for any codevector $\boldsymbol{c}' \neq \boldsymbol{c}$,

$$\mathsf{dist}\,(\boldsymbol{r}, \boldsymbol{c}') \ge n(\frac{1}{4} + \theta). \tag{5}$$

The decoding algorithm $A(c_0, \theta)$ accepts as input a received vector $\boldsymbol{r}$ and consists of the following steps.

1. Initialize the system to be in the zero-state $|\vec{0}\rangle|0\rangle$ where the first register has m qubits while the second one is a 1-qubit register.

2. Apply the transform W_2^m to the first register. Formally this corresponds to computing
$$(W_2^m \otimes I_2)(|\vec{0}\rangle|0\rangle) = W_2^m|\vec{0}\rangle I_2|0\rangle = 2^{-\frac{m}{2}} \sum_{\vec{v}\in \mathbf{Z}_2^m} |\vec{v}\rangle|0\rangle,$$
where I_2 denotes the 2-dimensional identity transform. Thus, the output of this step is the superposition $\sigma_1 := 2^{-\frac{m}{2}} \sum_{\vec{v}\in \mathbf{Z}_2^m} |\vec{v}\rangle|0\rangle$.

3. In quantum parallelism, store the v-th bit $\boldsymbol{r}(\vec{v})$ of the received word $\boldsymbol{r}$ in the second register, producing
$$\sigma_2 := 2^{-\frac{m}{2}} \sum_{\vec{v}\in \mathbf{Z}_2^m} |\vec{v}\rangle|\boldsymbol{r}(\vec{v})\rangle.$$
The unitariness of this computation has been discussed in many related works (see, e.g., [8]).

4. Apply the conditional sign-shift transform P_2 to the second register:
$$\sigma_3 := (I_2^m \otimes P_2)\sigma_3 = 2^{-\frac{m}{2}} \sum_{\vec{v}\in \mathbf{Z}_2^m} (-1)^{\boldsymbol{r}(\vec{v})}|\vec{v}\rangle|\boldsymbol{r}(\vec{v})\rangle.$$

5. Apply the inverse transform of Step (3) to the second register, producing
$$\sigma_4 := 2^{-\frac{m}{2}} \sum_{\vec{v}\in \mathbf{Z}_2^m} (-1)^{\boldsymbol{r}(\vec{v})}|\vec{v}\rangle|0\rangle.$$

6. Apply the transform W_2^m to the first register:
$$(W_2^m \otimes I_2)\sigma_4 = 2^{-m} \sum_{\vec{v}\in \mathbf{Z}_2^m} (-1)^{\boldsymbol{r}(\vec{v})} \sum_{\vec{u}\in (\mathbf{Z}_2^m)^*} (-1)^{<\vec{u},\vec{v}>}|\vec{u}\rangle|0\rangle = \sum_{\vec{u}\in (\mathbf{Z}_2^m)^*} \frac{\widehat{F}_{\boldsymbol{r}}(\vec{u})}{n}|\vec{u}\rangle|0\rangle,$$
where the last equality follows by Definition 2.1.

7. Measure the first register, obtaining a certain vector $\vec{u}$.

8. Repeat Steps (1)-(7) $t = c_0 \ln n$ times. Set the decoding result to the vector $\vec{u}$ obtained the greatest number of times by the measurements in Step (7).

Except for Steps 3 and 5, the amount of quantum bit operations performed in the computation is clearly $O(m) = O(\log n)$. For Step 3, we notice that given a received vector $\boldsymbol{r}$ as input, there exists a classical circuit of linear size that can implement this in time $O(\log n)$. Following a by now standard fact (see, e.g., §3 in [8]), this classical computation can be made reversible for only a constant factor cost in time. Therefore, Steps 3 and 5 also have time complexity $O(\log n)$ (although the quantum circuit implementing them is of size linear in n). Since these steps need to be repeated $t = O(\log n)$ times, the overall time complexity is $O(\log^2 n)$.

4 Probability of decoding error

Let $\boldsymbol{r} \in \mathbf{Z}_2^n$ be a vector. Suppose $\boldsymbol{r} \in D_\theta(C)$ and let $\vec{u}_{\boldsymbol{r}} \in (\mathbf{Z}_2^m)^*$ be the message that corresponds to the codevector $\delta_\theta(\boldsymbol{r})$. As above, we denote this codevector by $\boldsymbol{c}_{\vec{u}_{\boldsymbol{r}}}$. In this section we prove that the probability P_e that the algorithm fails to find this codevector given $\boldsymbol{r}$ as the input, behaves as n^{-c}.

We need Chernoff-type bounds on large deviations. The following technical theorem can be found, for instance, in [7], [1].

Theorem 4.1 *Let $X_1, X_2, \ldots, X_t$ be i.i.d. Bernoulli variables such that for $1 \le i \le t$, $Pr[X_i = 1] = p$, where $p \in (0,1)$. Let $X = \sum_{i=1}^t X_i$ and thus $\mu = \mathsf{E}[X] = pt$. Then*

1. *for any $0 < \delta < 1$, $Pr[X < (1-\delta)\mu] < e^{-\frac{\delta^2\mu}{2}}$;*
2. *for any $0 < \delta \le 2e-1$, $Pr[X > (1+\delta)\mu] < e^{-\frac{\delta^2\mu}{4}}$; and*
3. *for any $\delta > 2e-1$, $Pr[X > (1+\delta)\mu] < 2^{-\delta\mu}$.*

This enables us to prove the following result.

Theorem 4.2 *For any word $f \in \mathcal{D}$, Algorithm $A(c_0, \theta)$ fails to return $\vec{u}_f$ with probability $P_e < n^{-c}$, where*

$$c = \min\{8c_0\theta^6, \frac{4\theta^4 c_0}{(1-4\theta)^2} - 1, \frac{2\theta^2 c_0}{\log e} - 1, \} - 1.$$

Proof: The probability that $\vec{u}$ is obtained in a measurement in Step (7) of the algorithm equals $p_{\vec{u}} = \frac{|\widehat{F}_f(\vec{u})|^2}{n^2}$. Let $X_i(\vec{u})$, $1 \le i \le t$, be the Bernoulli random variables such that $X_i(\vec{u}) = 1$ according as $\vec{u}$ is the result of the measurement in iteration i or not. Let

$$X(\vec{u}) = \sum_{i=1}^{t} X_i(\vec{u}).$$

Then the expected number of times that $\vec{u}$ is obtained in t measurements is $\mathsf{E}[X(\vec{u})] = p_{\vec{u}}t$. We have

$$\begin{aligned} \mathsf{E}[X(\vec{u}_{\boldsymbol{r}})] - \mathsf{E}[X(\vec{u})] &= (p_{\vec{u}_{\boldsymbol{r}}} - p_{\vec{u}})t = \frac{t}{n^2}\left[|\widehat{F}_{\boldsymbol{r}}(\vec{u}_{\boldsymbol{r}})|^2 - |\widehat{F}_{\boldsymbol{r}}(\vec{u})|^2\right] \\ &= \frac{t}{n^2}(|\widehat{F}_{\boldsymbol{r}}(\vec{u}_{\boldsymbol{r}})| - |\widehat{F}_{\boldsymbol{r}}(\vec{u})|)(|\widehat{F}_{\boldsymbol{r}}(\vec{u}_{\boldsymbol{r}}) + |\widehat{F}_{\boldsymbol{r}}(\vec{u})|) \end{aligned} \quad (6)$$

Now, since $\boldsymbol{c}_{\vec{u}_{\boldsymbol{r}}}$ is the closest codevector to $\boldsymbol{r}$ and since the covering radius of C is $\frac{n}{2}$, we have

$$\widehat{F}_{\boldsymbol{r}}(\vec{u}_{\boldsymbol{r}}) = n - 2\mathsf{dist}\,(\boldsymbol{c}_{\vec{u}_{\boldsymbol{r}}}, \boldsymbol{r}) \ge 0.$$

Thus we need to consider two cases depending on the sign of $\widehat{F}_{\boldsymbol{r}}(\vec{u})$. However, (6) is symmetric w.r.t. this sign. Therefore, we obtain

$$\begin{aligned} \mathsf{E}[X(\vec{u}_{\boldsymbol{r}})] - \mathsf{E}[X(\vec{u})] &= \frac{4t}{n^2}(\mathsf{dist}\,(\boldsymbol{c}_{\vec{u}}, \boldsymbol{r}) - \mathsf{dist}\,(\boldsymbol{c}_{\vec{u}_{\boldsymbol{r}}}, \boldsymbol{r}))(n - \mathsf{dist}\,(\boldsymbol{c}_{\vec{u}_{\boldsymbol{r}}}, \boldsymbol{r}) - \mathsf{dist}\,(\boldsymbol{c}_{\vec{u}}, \boldsymbol{r})) \\ &\ge 4t\theta^2. \end{aligned} \quad (7)$$

by (1)-(2).

Let $\alpha = 2\theta^2$. Consider the following two events:

$$\mathcal{E}_1 : X(\vec{u}_r) \geq E[X(\vec{u}_r)] - \alpha t,$$
$$\mathcal{E}_2 : X(\vec{u}) \leq E[X(\vec{u})] + \alpha t \text{ for any } \vec{u} \neq \vec{u}_r$$

From (7) it is clear that $(\mathcal{E}_1 \cap \mathcal{E}_2) \subset$ [correct decoding]. Therefore,

$$P_e \leq \Pr[\bar{\mathcal{E}}_1 \cup \bar{\mathcal{E}}_2] \leq \Pr[\bar{\mathcal{E}}_1] + \Pr[\bar{\mathcal{E}}_2].$$

Thus to prove the theorem we need to show that the probability on the right-hand side is bounded above by n^{-c}. Note that from (4)-(5) we have

$$\begin{aligned} E[X(\vec{u}_r)] &\geq 4t\theta^2, \\ E[X(\vec{u})] &\leq t(\frac{1}{4} - 2\theta + 4\theta^2), \quad \vec{u} \neq \vec{u}_r. \end{aligned} \tag{8}$$

Now we are in a position to apply Theorem 4.1. Namely, by Part 1, taking into account the obvious $E[X(\vec{u}_r)] \leq t$, we have

$$\begin{aligned} \Pr[\bar{\mathcal{E}}_1] = \Pr[X(\vec{u}_r) < \mathsf{E}[X(\vec{u}_r)] - \alpha t] &< e^{-\alpha^2 \mathsf{E}[X(\vec{u}_r)]/2} \leq e^{-8t\theta^6} = n^{-8c_0\theta^6} \\ &< n^{-c}/2. \end{aligned}$$

Now suppose that $\vec{u} \neq \vec{u}_f$. We use Parts 2 and 3 of Theorem 4.1 together with (8):

$$\begin{aligned} Pr[X(\vec{u}) > E[X(\vec{u})] + \alpha t] &< \max\left\{\exp(\frac{-\alpha^2 t}{4(\frac{1}{4} - 2\theta + 4\theta^2)}), 2^{-\alpha t}\right\} \\ &= \max\{n^{-\frac{\alpha^2 c_0}{(1-4\theta)^2}}, n^{-\frac{\alpha c_0}{\log e}}\} \\ &< n^{-(c+1)}/2. \end{aligned}$$

Since there are at most n such $\vec{u}$ in question, we get $\Pr[\bar{\mathcal{E}}_2] < n^{-c}/2$. This completes the proof of the theorem. ∎

Remarks. 1. Note that the 1st order Reed-Muller code $\mathrm{RM}(1, m)$ equals $C \cup (\mathbf{1} + C)$, and therefore, the algorithm can be applied to this code as well. However, then we can only claim that decoding result is either the correct codevector or its complement.

2. Note that since any binary $[n, n-m]$ linear code can be embedded in the simplex code of length 2^m, our algorithm can be applied to the decoding of general linear codes. However, this does not seem to improve upon the application of the general search method [6].

3. In principle, the θ's in the definitions of D_θ (1) and E_θ (2) do not have to be the same. The only change that this would incur is in the expression for c in Theorem 4.2.

Acknowledgment. We would like to thank Lov Grover and Francis Zane for helpful discussions on quantum computing.

References

[1] N. Alon, J. H. Spencer, The probabilistic method. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., New York, 1992. xvi+254 pp.

[2] C. Bennett, E. Bernstein, G. Brassard, and U. Vazirani, Strengths and weaknesses of quantum computing, SIAM J. Comput. **26** (5) (1997), 1510–1523.

[3] E. Berlekamp, R. J. McEliece and H. C. A. van Tilborg, On the inherent intractability of certain coding problems, IEEE Trans. Inform Theory, **IT-29** (3) (1978), 384–386.

[4] M. Boyer, G. Brassard, P. Høyer, and A. Tapp, Tight bounds on quantum searching, Fortsch.Phys. 46 (1998) 493-506.

[5] R. R. Green, A serial orthogonal decoder, JPL Space Programs Summary, Vol. 37-39-IV (1966), pp. 247–253.

[6] Lov K. Grover, A fast quantum mechanical algorithm for database search, Proceedings, 28th Annual ACM Symposium on the Theory of Computing (STOC), May 1996, pp. 212-219.

[7] R. Motwani, P. Raghavan, Randomized algorithms. Cambridge University Press, Cambridge, 1995. xiv+476 pp.

[8] P. W. Shor, Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer, SIAM J. Comput. **26** (5) (1997), 1484 - 1509

[9] Special Section on Quantum Computation, SIAM J. Comput. **26** (5) (1997).

On Two-Dimensional Arrays and Crossword Puzzles

Jack Keil Wolf
Center for Magnetic Recording Research
Mail Code: 0401
University of California-San Diego
La Jolla, CA 92093-0401
jwolf@ucsd.edu

Paul Siegel
Dept. of Electrical & Computer Engineering
Mail Code: 0407
University of California-San Diego
La Jolla, CA 92093-0407
psiegel@ ucsd.edu

Abstract.

We consider large two-dimensional arrays where each row and each column must satisfy certain constraints. We relate this problem to an assertion made by Shannon regarding the existence of large two-dimensional crossword puzzles.

One-dimensional constrained sequences.

We begin by considering an alphabet containing M symbols and one-dimensional sequences of these symbols that satisfy a set of well defined constraints. To narrow the focus of this paper we will concentrate on a special class of such one-dimensional constrained sequences which consist of concatenations of isolated words from some dictionary. Details of this special class follow.

We first assume that one of the M symbols is of special significance and is called the space symbol. The collection of the remaining (M-1) symbols will be called non-space symbols. We assume that we have a dictionary of Q allowable words. The words in this dictionary are assumed to be sequences of the (M-1) non-space symbols. We assume that the i-th word of the dictionary is of length L_i and that all the words in the dictionary are distinct. Finally we assume that the constraints allow for every concatenation of words from the dictionary provided that consecutive words are separated by one (or in some cases more than one) space symbol.

One example of such a constrained system would be the sequences of English words in the rows or columns of a crossword puzzle as contained in many American newspapers or magazines. Here M=27, the black space between words is considered the space symbol and there are 26 non-space symbols. In this case, we allow for one or more than one space symbol between words in the dictionary to allow for one or more than one black square between consecutive words.

A second example of such a constrained system would be binary (d,k) sequences. The alphabet consists of the two symbols 0 and 1. The usual description of such binary sequences is that the symbols must satisfy the following two constraints.

(1) d-constraint: Two consecutive 1's must be separated by a run of at least d consecutive 0's.

(2) k-constraint: The maximum length of a run of consecutive 0's is k.

Note, however, the following equivalent description of such (d,k) binary sequences which applies when $d \geq 1$. The M=2 binary symbols are the space symbol "1" and the non-space symbol "0". The dictionary contains the Q=(k-d+1) code words consisting of runs of j 0's where j takes on the (k-d+1) values from d to k. Furthermore, every word in the dictionary can be followed by exactly one space symbol (i.e., one "1"). The modifications to this description that must be made for the case of d=0 are straightforward.

If any concatenation of the M symbols is allowable, there are exactly M^n distinct sequences of length n. It is desirable to know how many distinct sequences of length n can be formed for the constrained system. Shannon [1] provided us with several methods of computing this quantity. One method is very easy to understand in terms of our dictionary description of these constrained sequences.

Assume first the case where there is one and only one space between adjacent code words. Let N(n) be the number of distinct sequences of length n in our constrained system. Note that any sequence of length n can be thought of as a sequence of length (n-j-1) followed by a space followed by a code word of length $j \geq 1$. Then, if N(n-j-1) is the number of distinct sequences of length (n-j-1) and if A_j denotes the number of distinct words in our dictionary of length j, then the number of distinct sequences of length N(n) would equal the sum over all code word lengths j of the product of A_j and N(n-j-1).

From the standard theory of linear difference equations one knows that the solution for N(n) is a sum of exponentials of the form γ^n which for large n is dominated by the largest real root of an algebraic equation found by substituting γ^j for N(j) in the original linear difference equation.

For the case of binary (d,k) codes, the value of the limit as n approaches infinity of $\log_2[N(n)]/n$ has been computed by many authors. We will denote this quantity by $C_1(d,k)$. $C_1(d,k)$ has been referred to as the capacity (or the maximum entropy) of the constrained system. For our purposes, we need only know that $C_1(d,k)$ can be computed for all values of d and k and that the value of $C_1(d,k)$ is non-zero for all d and k provided that $k > d$. One more fact that will be useful later is that $C_1(1,2)= C_1(2,4)$.

For the case of an arbitrary concatenation of English words as in the rows and columns of a crossword puzzle, one must take into account the fact that one can have more than one space between words. In that case, the above-mentioned difference equation is modified by adding the additional term N(n-1) to the sum of products of A_j and N(n-j-1). The method of solution, however is identical to that described previously.

For this case, the value of the limit as n approaches infinity of $\log_2[N(n)]/n$ has been computed by Gilbert [2] by counting the number of English words of each length in two different English dictionaries. Calling this quantity C_1(English), Gilbert computed C_1(English) to be 2.444 bits/letter for a dictionary of size 34,897 and computed C_1(English) to be 2.782

bits/letter for a dictionary of size 233,614. Again, C_1(English) has been referred to as the capacity or the maximum entropy of the constrained system.

Two-dimensional constrained arrays and crossword puzzles.

We next consider two-dimensional arrays of symbols from an alphabet of size M where the arrays are such that every row and every column must individually satisfy the previously described constraint for one-dimensional sequences. That is, every row and every column of the array must be a concatenation of words from a dictionary where adjacent words are separated by one or possibly more than one space. Two examples of such arrays which are the two-dimensional extension of the one-dimensional examples are:

1. Two-dimensional crossword puzzles as found in many American newspapers or magazines where every row and every column consists of a sequence of words from an English dictionary and such that every pair of consecutive words in every row and every column are separated by one or more spaces (i.e., black squares).

2. Binary two-dimensional (d,k) arrays where each element of the array is a "0" or a "1" and such that every row and every column satisfies the previously specified d-constraint and k-constraint.

For a specified constraint we are interested in how many such distinct arrays can be constructed for an array with m rows and n columns. We call this number N(m,n).

It is clear that for large m and n, for every constraint for which the one-dimensional capacity is non-zero, N(m,n) must grow at least exponentially with m or with n. This is the case since if one chooses the first column (or row) to satisfy the one-dimensional constraint, one can always have the rows (or columns) satisfy the constraint by choosing each subsequent column (or row) as a shift by one symbol up or down (or to the left or right) of the previous column (or row).

However, one might suspect that in some cases, that for large m and n, N(m,n) might grow exponentially with the product of m and n. Indeed let us define the two-dimensional capacity or maximum entropy of the array as the limit as m approaches infinity and as n approaches infinity of $\log_2[N(m,n)]/mn$. For the case of two-dimensional crossword puzzles of English words we call this quantity C_2(English), while for the case of two-dimensional binary (d,k) arrays we call this quantity $C_2(d,k)$.

It is of specific interest to us that Shannon discussed two-dimensional crossword puzzles of English letters in his 1948 paper [1]. In part, Shannon stated:

> "The ratio of the entropy of a source to the maximum value it could have while still restricted to the same symbols will be called its *relative entropy*."
>
> "One minus the relative entropy is the *redundancy*."

> "The redundancy of a language is related to the existence of crossword puzzles. If the redundancy is zero any sequence of letters is a reasonable text in the language and any two dimensional array of letters forms a crossword puzzle. If the redundancy is too high the language imposes too many constraints for large crossword puzzles to be possible. A more detailed analysis shows that if we assume the constraints imposed by the language are of a rather chaotic and random nature, large crossword puzzles are just possible when the redundancy is 50%. If the redundancy is 33%, three dimensional crossword puzzles should be possible, etc."

No further explanation of these remarks has been found in the subsequent publications of Shannon nor in the work of other authors. As a result of one of the authors having presented some of this material at the Shannon Day Symposium at Bell Laboratories on May 18, 1998, a letter and other correspondence [2] was received from Edgar Gilbert reporting on some previous work that he had done on the subject and on his recollection of some discussions he had had with Shannon related to this work. The following argument was completed before having received the correspondence from Gilbert but seems to be corroborated by its contents.

In order to proceed we will make some conjectures as to how to interpret Shannon's statements. We first conjecture that Shannon's words "large crossword puzzles" are "possible" should be interpreted to mean that the two-dimensional capacity of the system is strictly greater than 0. Conversely, when Shannon states that, if the redundancy is too high, large crossword puzzles are not possible, we assume that Shannon meant that in that case the two-dimensional capacity is equal to 0.

We go on to conjecture that Shannon's description of the constraints as being of a "chaotic and random nature" refers to the fact that the joint statistics that govern the entries of the array are such that all of the rows and all of the columns are statistically independent of one another.

We now give a plausibility argument for Shannon's conclusion based upon these conjectures. We first note that the number of distinct arrays having m rows and n columns with elements from an alphabet with M symbols is equal to M^{mn}. For crossword puzzles made up of English words, M=27. There are N(n) distinct sequences of length n (or N(m) sequences of length m) that satisfy the constraint. For any of the M^{mn} arrays previously constructed, the probability that any row satisfies the constraint is $N(n)/M^n$, and the probability that any column satisfies the constraint is $N(m)/M^m$. Invoking our conjecture as to the meaning of "chaotic and random", the probability that every row and every column of the array satisfies the constraints is $[N(m)]^n[N(n)]^m/M^{2mn}$. Under this assumption, the average number of arrays for which all of the rows and all of the columns satisfy the constraints is equal to $[N(m)]^n[N(n)]^m/M^{mn}$.

Assuming that we take m and n very large, and denote by C the appropriate capacity or maximum entropy, we can use the approximation $N(m)=2^{mC}$ and $N(n)=2^{nC}$. Then writing $M=2^{\log(M)}$ (where the logarithm is taken base 2), we have that the average number of arrays having all of their rows and all of their columns satisfying the constraints is equal to $2^{mn(2C-\log(M))}$. Thus if 2C-log(M) > 0 (or equivalently C/log(M) > 0.5), the average number of arrays that have all rows and all columns satisfying the constraints grows exponentially with the product mn. A similar argument follows for higher dimensional arrays. This agrees with Shannon's conclusions.

We can improve on this result by first choosing arrays that satisfy only the constraints on the rows. Using the above argument, the average number of arrays that have all of their m rows satisfying the constraints is equal to $[N(n)]^m$. These arrays will have the symbols in each column occurring in accordance with their first order probability. Let H1 be the entropy corresponding to their first order probability. The probability that any column in these surviving arrays satisfies the constraints is then for large m approximately given by $N(m)/2^{mH1}$. The probability that all n columns of these surviving arrays satisfies the constraints is then given by $[N(m)]^n/2^{mnH1}$. Thus the average number of arrays that will have both their rows and columns satisfying the constraints is given by $2^{mn(2C-H1)}$. Thus if 2C-H1 > 0 (or equivalently C/H1 > 0.5), the average number of arrays that have all rows and all columns satisfying the constraints grows exponentially with the product mn. Again a similar argument follows for higher dimensional arrays. Specifically, this argument says that large three-dimensional crossword puzzles exist provided that C/H1 > 0.666.

Using the definition of redundancy as stated by Shannon and Gilbert's estimate of the maximum entropy of English words for the 233,614 word dictionary, one finds that the redundancy of English words is about 0.415. Following Shannon, this would allow for the construction of large two-dimensional crossword puzzles but not for large three-dimensional crossword puzzles.

Ignoring the allowed multiplicity of spaces between words, the first order entropy of English is known to be approximately 4.03 bits per symbol. Using Gilbert's estimate of the entropy of English words for the 233,614 word dictionary one finds that C/H1 = 0.690. Thus, large three-dimensional crossword puzzles of English words indeed may exist.

Two-dimensional (d,k) arrays.

We are on somewhat shaky ground trying to apply the previous ideas to binary (d,k) arrays since these constraints are far from being of a chaotic and random nature. However, it is of some interest to see what Shannon's ideas predict in that case.

Very little is known about the two-dimensional capacity, $C_2(d,k)$ of binary (d,k) arrays. In fact the exact value of capacity, $C_2(d,k)$, has only been computed for the case when $C_2(d,k)=1$ or $C_2(d,k)=0$. The former case is the trivial case of unconstrained arrays (i.e., d=0 and k=infinity). The latter case is known to occur [3] if and only if $d \geq 1$ and k=d+1. The case of $C_2(d,k) = 0$ for $d \geq 1$ and k=d+1 is predicted by Shannon's theory since $C_1(1,2) = 0.4057$ and $C_1(d,d+1) \leq .4057$ for d > 1. Thus for these cases, the redundancy is greater than 0.5 so one would not expect to be able to construct large two-dimensional arrays satisfying these constraints.

As stated previously, $C_1(2,4)= C_1(1,2)$ but also as stated previously $C_2(1,2)=0$ and $C_2(2,4) > 0$. Thus it is clear that Shannon's argument cannot be applied since both constraints have the same redundancy but one has a two-dimensional capacity equal to 0 and the other has a non-zero two-dimensional capacity. Our hope was that the modified argument might predict the correct behavior by taking into account the fact that the first order statistics for the two constraints are not the same. Indeed the (2,4) constraint has a higher value for C/H1 than does the (1,2) constraint but unfortunately both have C/H1< 0.5. Thus our initial skepticism in applying these ideas to this constraint was correct.

Concluding comments.

It is clear that there is much to be learned regarding two-dimensional constrained arrays. One possible avenue for future research is to attempt to improve further on the argument given related to the existence of large two-dimensional arrays. In particular, it is clear that the rows and columns of binary two-dimensional (d,k) arrays are not statistically independent so that one might attempt to model the dependence in some manner.

Acknowledgments.

The authors are indebted to Edgar Gilbert for his many contributions and to Kevin Shaughnessy who contributed to the initial understanding of Shannon's statements. This work was partially supported by the Center for Magnetic Recording Research at the University of California, San Diego and by the National Science Foundation under grants NCR-9405008 and NCR-9612802.

References

[1 C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 10, pp. 379-423, 623-656, October, 1948.

[2] E.N. Gilbert, private communication, 1998.

[3] A. Kato and K. Zeger, "On the capacity of two-dimensional run-length limited codes," *Proceedings of the 1998 International Symposium on Information Theory*, pg. 320, Cambridge, MA, August, 1998.

Coding for Channels with Local Data Dependent Noise*

Rolf Erstad
Provida A/S
Bergen
Norway
rolfe@ii.uib.no

Øyvind Ytrehus
Dept. of Informatics
University of Bergen
N-5020 Bergen, Norway
oyvind@ii.uib.no

September, 1998

Abstract

In many channels, the noise distribution depends on the transmitted or recorded data. We discuss properties of codes for such channels for such channels, and present some examples.

1 Introduction

In many recording and transmission systems, the probability distribution of the noise which corrupts a given channel symbol depends on the value of that symbol as well as on the value of a finite number of preceding and following symbols. Our starting point and motivating example for this study is derived from magnetic recording.

Consider a magnetic recording disk system where binary symbols are recorded by changing the direction of magnetization in a magnetic material. A sequence of symbols is recorded along a track in the disk, such that a fixed portion of the track (a *bitcell*) is assigned to each symbol. A "1" is written as a change in magnetization direction in the corresponding bitcell, while a "0" is recorded by simply not changing the direction of magnetization.

In order to read back the data, a read head passes along the track. Any change in magnetic flux causes an electric pulse at the output of the read head. Ideally, we can then associate a pulse with a recorded "1" in the bitcell while the absence of a pulse in a given bitcell will be interpreted as a recorded "0".

In a modern recording system, the data are recorded at so high density that pulses from adjacent or neighboring "1"s interact, this phenomenon is denoted Intersymbol Interference, or ISI. In order to simplify the process, it is common to equalize the channel output into a manageable ISI, defined by a partial response polynomial. Moreover, the retrieval process is complicated by two different noise processes:

- Additive White Gaussian Noise (AWGN) from the read head. Although the noise is white, i. e. different noise symbols are uncorrelated, the equalization process

---

*This research was supported by NFR Grants 107542/410, 107623/420, and 119390/431.

will color it, complicating the process of performing maximum likelihood sequence estimation or maximum likelihood decoding.

- Media Noise. This is due to imperfections in the magnetic material so that the magnetic transition is not written exactly and precisely where it is expected to occur. One consequence is that it is hard to predict the ISI from one pulse onto neighboring pulses. Note that this type of noise is data dependent since it occurs only if the channel symbol is nonzero.

A simple recording system model ignores the nonlinearities and colors of the noise processes and assumes that the noise is AWGN. In such a model, a maximum likelihood decoding is carried out simply by letting a Viterbi decoder (or detector) find the input sequence which is closest to the channel output sequence in terms of squared Euclidean distance. The squared Euclidean distance is an additive metric, and the distance between the received sequence and the noiseless output can be calculated symbol by symbol.

In a real magnetic recording system, we know that the noise is correlated, and Euclidean distance is not capable of using the information contained in the correlation. In this paper we study properties of systems where the data dependencies and the coloring of the noise is taken into account. In particular we will always assume that the data are encoded by an error correcting code, and we will study properties of such codes that simplify the decoding. (This also encompasses the trivial but common case where the data are basically uncoded.)

2 Local data dependent noise

The value of a sampled symbol is dependent on both the future symbols and the past symbols when correlated noise and ISI are present. Ideally, the sampled value depends on the length of the pulse response. Maximum Likelihood (ML) detection can be performed by a 2^{2L+1} state trellis, where $2L+1$ denotes the total length of the ISI [3]. Since media noise and ISI are dependent on the input sequence to the channel, the channel output will depend on the L previous and the L future noiseless channel symbols, in addition to the present symbol. The main distortion of the signal is due to media noise and ISI from the m nearest symbols to each side. A sub-optimal detection can be obtained by studying the m_p previous and the m_f future symbols in addition to the present one, and neglecting the influence from symbols "far" from the present one.

We will call the noise *local data dependent* if the noise affecting a particular symbol depends on m_p previous and m_f future noiseless output symbols.

Let $(x_1, \cdots, x_N)$ be the noiseless channel output sequence, and $(y_1, \cdots, y_N)$ the corresponding sampled and discretized channel output. A channel where the sampled output is corrupted by local data dependent noise is denoted a Discrete Local Data Dependent Channel (DLDDC) if

$$p(y_i|x_1, \cdots, x_N, y_1, \cdots, y_{i-1}) = p(y_i|\mathrm{x}_{i-m_p}^{i+m_f}, \mathrm{y}_1^{i-1}),$$

where $x_{i-m_p}^{i+m_f} = (x_{i-m_p}, \cdots, x_{i+m_f})$.

Further, since the noiseless channel symbols x_i usually belong to a relatively small alphabet while the output samples (even when discretized) will usually belong to a much larger alphabet, we will consider the simplified channel model where the probability distribution for the i-th symbol essentially depends on the channel symbol sequence viewed through a narrow window, i. e.,

$$p(y_i|x_1, \cdots, x_N, y_1, \cdots, y_{i-1}) = p(y_i|\mathbf{x}_{i-m_p}^{i+m_f}).$$

3 Viterbi decoder structures

We have carried out experiments with the micro-track model, developed by J. Caroselli and J. K. Wolf [2] a software model which includes the two noise processes described above. Our experiments indicate that a magnetic recording channel equalized to a partial response channel is modeled well as a DLDDC channel, with the parameters m_p and m_f depending on the actual system parameters. These results must be regarded as snapshots into the world of recording systems, valid only for the specific (but typical) set of parameters that we used.

If we want to design a decoder for a DLDDC channel, there are several approaches to follow.

3.1 A block oriented, inaccurate approach

Let us assume that we have a trellis where the branch between two states at successive time instants is associated with a label consisting of n channel symbols. For each state in the trellis the decoder makes a local decoding decision, determining which of the paths leading into that state is more likely. For a DLDDC this is not optimal. If $m_p > 0$, then potentially a path that was discarded at a previous state could turn out to be the best in hindsight. Similarly, if $m_f > 0$, then symbols on future branches in the trellis influence the calculation of current metric values, but those future symbols are in general unknown.

One way to deal with this problem is to assign metric values to the branch, simply averaging over all possible future symbols. For the preceding symbols, we can either again take the average over all possible channel symbols, or better (?), pick the last symbols from the surviving path into the previous state.

We have done this for a typical recording channel, using the microtrack model. The results are shown in Figure 1. Details can be found in [1]. Observe from the figure that a considerable gain gan be achieved by using this decoder metric in place of the common Euclidean metric. This motivates a further study of how maximum likelihood decoding will perform. In this paper we focus on the properties of a trellis to be used for maximum likelihood decoding on a DLDDC channel.

3.2 A symbol oriented decoder metric

In order to be able to design a maximum likelihood decoder, we want to use a decoder metric (abusing notation; strictly speaking this is a distance measurement and not a metric) which

- is independent of the coding,
- is symbol oriented (the distance should be decided on a symbol by symbol basis),

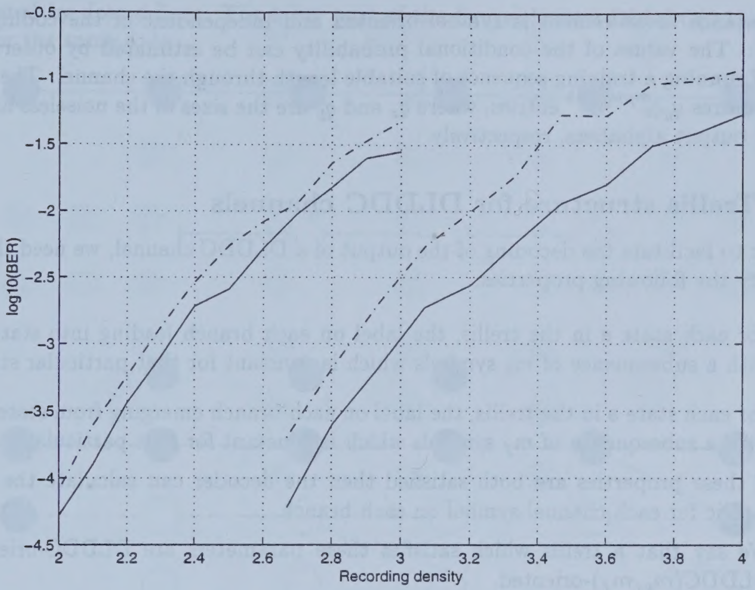


Figure 1: *Performance for the rate 8/9(left) and rate 2/3(right) codes at S-AWGN-R=20 dB, equalized to a partial response channel with transfer polynomial* $1 - D^2$ *(class IV). The curves show physical recording density, i.e, they are not rate-normalized. Solid lines: New decoder metric, dashed lines: Euclidean metric.*

- requires only short training sequences, and
- that can be implemented by a look-up table of moderate size.

Let $\mathbf{y}$ be a real channel output sequence, and $\mathbf{x} \in C$ be the noiseless output sequence of a channel coded with a code C. An optimal decoding scheme will maximize $p(\mathbf{x}|\mathbf{y})$. A maximum likelihood decoder will decode to the sequence $\mathbf{x}$ which maximizes $p(\mathbf{y}|\mathbf{x})$. For sequences $\mathbf{x}$, $\mathbf{y}$ of length N on a discrete memoryless channel we have

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} p(y_i|x_i).$$

However, since a given output symbol, y_i, on a DLDDC channel depends on the noiseless output segment $\mathbf{x}_{i-m_p}^{i+m_f}$ we can use:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_{i-m_p}^{i+m_f}),$$

where we assume that x_i is known for $x_{-m_p+1}, \cdots, x_1, \cdots, x_N, x_{N+1} \cdots, x_{N+m_f}$.

For application in a Viterbi algorithm, we prefer an additive distance measure, and use:

$$f(y_i|\mathbf{x}_{i-m_p}^{i+m_f}) = -\log p(y_i|\mathbf{x}_{i-m_p}^{i+m_f}).$$

This distance measurement is symbol oriented and independent of the coding of the channel. The values of the conditional probability can be estimated by observing the effect of running a training sequence of suitable length through the channel. The look-up table requires $q_y q_x^{m_f+m_p+1}$ entries, where q_x and q_y are the sizes of the noiseless and noisy channel output alphabets, respectively.

3.3 Trellis structure for DLDDC channels

In order to facilitate the decoding of the output of a DLDDC channel, we need the trellis to satisfy the following properties:

- For each state s in the trellis, the label on each branch leading into state s ends with a subsequence of m_p symbols which is constant for that particular state s.
- For each state s in the trellis, the label on each branch emerging from state s starts with a subsequence of m_f symbols which is constant for that particular state s.

  If these properties are both satisfied then the decoder can calculate the decoder metric for each channel symbol on each branch.

  We say that a trellis which satisfies these parameters are DLDDC-oriented, or DLDDC(m_p, m_f)-oriented.

4 DLDDC-oriented trellises

We have experimented further with the micro-track model. After developing a distance measure for the new decoder metric, we found that a code with good minimum squared Euclidean distance will often have good distance properties also with the new distance [1]. We also know that for many channels, high rate codes with good minimum squared Euclidean distance are found among the codes with good Hamming distance.

In the following we will therefore mainly focus on how we can modify well known codes with a good minimum Hamming distance in such a way that their trellises become DLDDC-oriented. We will consider both block and convolutional codes.

An encoder is defined as minimal if it has the shortest constraint length among the encoders defining the same code. From this definition it also follows that a minimal encoder will have fewer states than an non-minimal encoder. For a more formal definition of minimal encoders, see [4].

Lacking a general channel model, we will for simplicity assume a binary channel alphabet, unless explicitly stated otherwise. Note that this deviates from our origin in a magnetic recording model.

Let us start out with a code whose minimal trellis is non-DLDDC oriented. In order to create a trellis which is DLDDC(0,1) oriented, we start with the old, minimal trellis. Then we split every state in the old graph into two copies, one out of which all the branches are labeled by a leading "0", and the other with leading "1"s. We also duplicate the branches leading into the old states. In general this may lead to a trellis that is non-minimal. That is, there will usually exist trellis representations with fewer states for the same code. A state in the new trellis may have more than one emerging branch, but each of these will have the same leading m_f output symbols.

The upper part of Figure 2 depicts a traditional minimal binary trellis for a $[5, 4, 2]$ parity check block code. In the figure, "0"s are represented by whole lines, while "1"s are represented by dotted lines. The lower part of the figure shows a DLDDC(0,1)-oriented trellis for the same code.

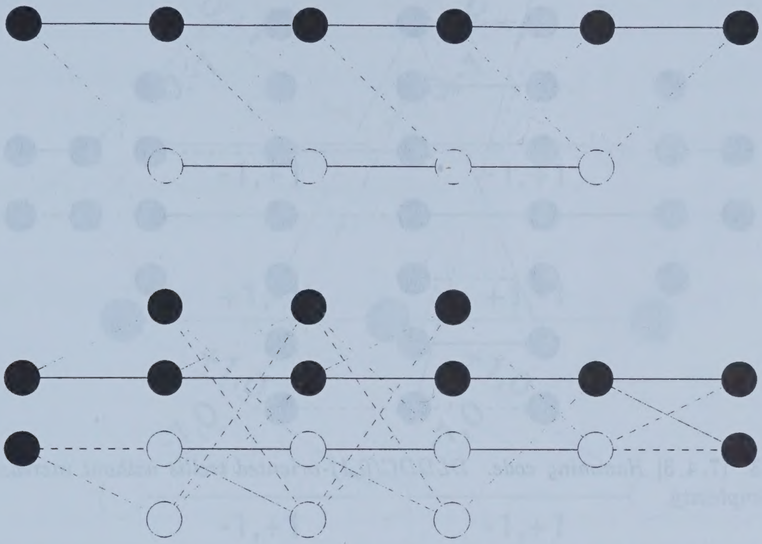


Figure 2: *Trellises for* $[5,4,2]$ *parity check code. Upper part shows a traditional bit oriented trellis, while the lower part shows a DLDDC(0,1)-oriented trellis of the same code.*

Figure 3 shows a DLDDC(0,1)-oriented trellis binary trellis for a $[7,4,3]$ Hamming code. The maximum number of states at any time instant of this trellis is 8. It is well known that any trellis for this code requires at least 8 states in at least one of the time instants, so this example shows that a DLDDC(0,1)-oriented trellis is not necessarily very much more complex than an ordinary trellis.

Figure 4 shows a Manchester (or FM) code that used to be popular in magnetic recording systems. Here the code is applied to a simple precoded partial response channel with transfer polynomial $1-D$. Hence, in this example, the noiseless channel output is ternary, with values in $\{-1,0,1\}$. We see from Figure 4 that the trellis is naturally either DLDDC(1,0)-oriented or DLDDC(0,1)-oriented, depending on the order of the symbols.

It is possible to characterize the parity check matrix and the generator matrix of a binary convolutional code with a natural minimal DLDDC(0,1)-oriented trellis.

Let $H(D)$ be a polynomial $(n-k)\times n$ parity check matrix for a convolutional code. Then denote by H_0 the $(n-k)\times n$ matrix consisting of the constant terms of the entries of $H(D)$.

Theorem 1. *A binary convolutional code has a minimal DLDDC(*$0,m_f$*)-oriented trellis if and only if there are at least* m_f *rows of* H_0 *such that (i) all the nonzero entries of these* m_f *rows are within the first* m_f *columns, and (ii) the* $m_f\times m_f$ *square submatrix consisting of the first* m_f *columns of these* m_f *rows has full rank.*

Sketch of proof: If: Assume that the parity check matrix has rows as specified above. In this case it has m_f syndrome equations where the solution depends on the current state and the m_f first symbols on the outgoing edges, i.e. the solution of the set is unique, and the m_f first symbols are fixed. **Only if:** Assume that the parity

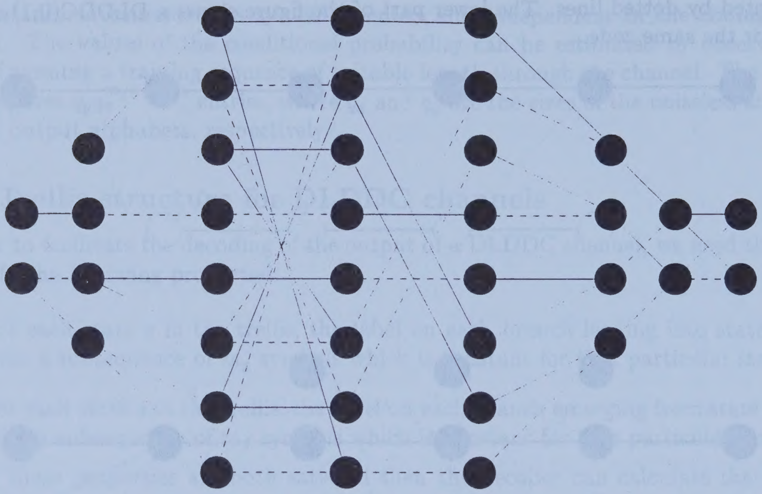


Figure 3: $[7,4,3]$ *Hamming code. DLDDC(0,1)-oriented trellis without increased total state complexity.*

check matrix is not on the form described above. Let l be the largest number of linear independent rows such that all its constant terms are within the m_f first columns. If $l < m_f$ we only have l syndrome equations with m_f unknowns, and the solution is not unique.

Q.E.D.

Alternatively we can study the generator matrix:

Theorem 2. *A binary convolutional code is described by a minimal DLDDC(0,m_f)-oriented trellis if its generator matrix does not have any constant term in the first m_f columns.*

Sketch of proof: **If:** Assume that the first m_f columns of the generator matrix do not have constant terms. Then the m_f first symbols are given by the encoder state only. **Only if:** If the generator matrix has a constant term different from zero in one of its first m_f column, say position (i, j), then the output symbol in position j will depend on both the encoder state and the input symbol in position i. That is, this symbol will assume two different values in two branches emanating from a common state.

Q.E.D.

We can make similar statements about the DLDDC(m_p, m_f)-oriented trellises with nonzero m_p, but we omit the details here. In particular, by expanding the set of states in a similar fashion as in the examples above, we need at most four times as many states in a DLDDC(1,1)-oriented trellis as what we had in the original, minimal trellis. Sometimes it is possible to do better, however.

Theorem 3. *Consider a binary convolutional code of rate $(n-1)/n$, with a polynomial parity check matrix $H(D) = (h_1(D), ..., h_n(D))$. Then the minimal trellis (with 2^ν*

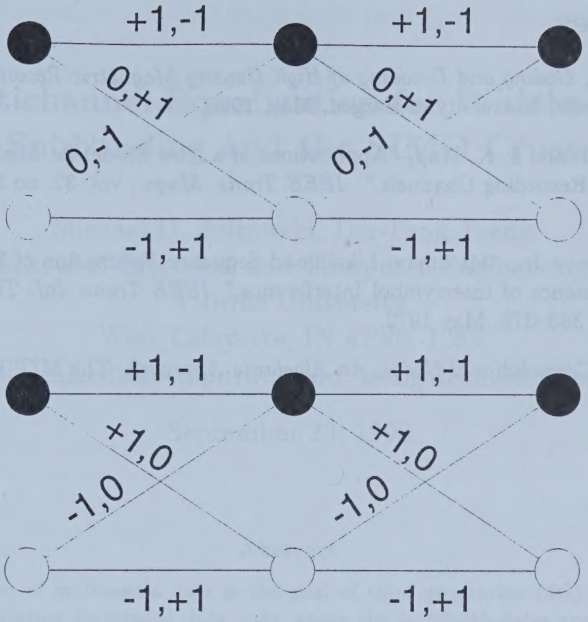


Figure 4: *Manchester code. The upper part pictures the traditional representation, which is DLDDC(1,0)-oriented, while the lower part is a DLDDC(0,1)-oriented trellis obtained by reversing the branch labels.*

states) for the code is DLDDC(1,1)-oriented if and only if the only constant term in any of the polynomials occurs in $h_1(D)$*, and the only term of degree* ν *occurs in* $h_n(D)$.

The proof is similar to that of Theorem 1. In particular, we have

Theorem 4. *Consider any binary convolutional code of rate 1/2, parity check matrix* $H_I(D) = (h_1(D), h_2(D))$, *and complexity* ν, *where* $\nu = \deg h_1(D) \geq \deg h_2(D)$. *Then the minimal trellis of the code of rate 1/2, parity check matrix*

$$H_{II}(D) = \left(h_1(D), D^{\deg h_1(D) - \deg h_2(D) + 1} h_2(D)\right),$$

and complexity $\nu + 1$, *is DLDDC(1,1)-oriented.*

Note that the two codes in Theorem 4 have the same free Hamming distance. For $n > 2$, in order to modify the parity check matrix in a similar way we require that all the polynomials $h_2(D), \ldots, h_{n-1}(D)$ have degree less than ν. However, it is still sometimes possible to find codes of rate $(n-1)/n$ whose minimal trellis is DLDDC(1,1)-oriented, such that its complexity is only one more than the simplest $(n-1)/n$ code with the same free distance. The table below shows examples of this.

Rate	ν	$H(D)$	d_{free}
2/3	5	$(D^4 + D^3 + D^2 + D + 1, D^4 + D, D^5 + D^4 + D^3 + D)$	5
2/3	8	$(D^7 + D^6 + D^5 + D^4 + D^3 + D + 1,$ $D^7 + D^2 + D, D^8 + D^6 + D^5 + D^4 + D^2)$	8
3/4	4	$(D^3 + D^2 + 1, D, D^3 + D^2 + D, D^4 + D^2 + D)$	4

References

[1] R. Erstad, *Coding and Decoding of High Density Magnetric Recording Systems*, Dr. Scient. thesis, University of Bergen, May, 1998.

[2] J. Caroselli and J. K. Wolf, "Applications of a New Model for Media Noise Limited Magnetic Recording Channels," *IEEE Trans. Magn.*, vol. 32, no 5, pp. 3917–3919, Sep. 1996.

[3] G. D. Forney Jr., "Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 3, pp. 363–378, May 1972.

[4] P. Piret, *Convolutional Codes, An Algebraic Approach,* The MIT Press, 1988.

Blind Multichannel Equalization for Wideband TDMA Based on Subbanding and the SIMO Cross-Relation [†]

Michael D. Zoltowski, Der-Feng Tseng
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285
e-mail: mikedz@ecn.purdue.edu, tseng@ecn.purdue.edu

September 23, 1998

Abstract

Transmission of multimedia data is the goal of third generation (3G) cellular radio systems necessitating high-speed data links where the multipath delay spread generally encompasses many symbol intervals. The channel impulse response for each antenna in a prototypical high-speed data link is many symbol intervals in duration and generally sparse. In order to reduce the number of parameters that need to be estimated in order to characterize the channel, researchers have recently proposed a number of schemes for direct estimation of the multipath parameters. However, this approach requires a relatively long observation interval in order to accurately estimate the complex gain of each dominant multipath at each antenna. We here propose an alternative approach based on subbanding and the Cross-Relation Method (CRM) of Xu, Liu, Tong, and Kailath. Basically, we apply the CRM in each subband using basis functions derived from the symbol waveform and the bandpass filter to blindly identify the frequency response of each channel (antenna) in the given subband, and then perform zero-forcing equalization across antennas for each subband. A method is presented for proper phasing of each subband channel estimate in order to reconstruct the original transmitted data signal. Residual multipath effects are then removed via a fractionally-spaced DFE the taps of which only encompass the effective duration of the original transmitted symbol, which is a fraction of the multipath delay spread. Simulations are presented demonstrating accurate identification of multiple channels under adverse multipath conditions, moderate SNR, and with a relatively small number of symbols.

[†] This research was supported by the Air Force Office of Scientific Research under grant no. F49620-97-1-0275, the Army Research Office's Focused Research Initiative under grant number DAAHO4-95-1-0246, and the National Science Foundation under grant no. MIPS-9708309.

1 Introduction: Subband Model

Employing linear modulation with a finite alphabet in a wireless digital communication system, the received signal at the i-th antenna, $i = 1, ..., N_A$, may be expressed as

$$y_i(t) = \sum_{m=0}^{N_b-1} b[m]h_i(t - mT_o) + n_i(t), \quad i = 1, ..., N_A, \tag{1}$$

where $b[m]$ is the m-th transmitted symbol, N_b is the number of symbols comprising the burst, $1/T_o$ is the symbol rate, and $n_i(t)$ is the noise process at the i-th antenna after front-end matched filtering with $p_{sr}(-t)$, the suboptimal demodulator, that matches the transmit pulse symbol waveform, $p_{sr}(t)$, having a square-root raised cosine spectrum, for example.

The "channel" impulse response associated with the i-th antenna, $h_i(t)$, may be expressed as $h_i(t) = p_{rc}(t) * h_{RF_i}(t)$, where "$*$" denotes linear convolution and $p_{rc}(t) = p_{sr}(t) * p_{sr}(-t)$ has a raised cosine spectrum and is referred to as the Nyquist symbol waveform with a roll-off factor β. $h_{RF_i}(t)$ is the RF multipath propagation channel that is approximated by a sum of Q dominant specular multipaths as $h_{RF_i}(t) = \sum_{k=1}^{Q} g_{ik}\delta_a(t - \tau_k)$, where g_{ik} and τ_k are the complex gain and delay, respectively, of the k-th multipath arrival at antenna i, and $\delta_a(t)$ is the Dirac delta function. An upper bound on the delay spread is assumed known and without loss of generality selected to be an integer multiple of $T_o/2$: $\max\{\tau_k\} < \tau_{max} = JT_o/2$. Note that we are dealing with the wideband case where the value of J may be large.

For equalization purposes, it is only necessary to accurately determine the frequency response of the RF multipath channel associated with each antenna over the signal bandwidth. In a narrowband setting, we showed in [1] that this may be achieved by employing a small number of delays equi-spaced across the delay spread, regardless of the actual multipath delay times. In order to utilize the techniques developed in [1] here for the wideband case, consider passing the signal received at each antenna through a bank of M analysis filters, as shown in Figure 1, where each filter passes a different narrow subband. We initially consider a uniform filter bank wherein the passband of each filter is the same. For reasons to presented later, it is important that adjacent subbands have significant overlap. Although the subbanding will be done via digital filtering after A/D conversion at twice the symbol rate, it will be useful at times for conceptual reasons to also view the subbanding operation in continuous time wherein the impulse response of the ℓ-th bandpass filter is denoted $f_\ell(t)$, $\ell = 1, ..., M$.

The output of the ℓ-th bandpass filter, $y_{i,\ell}(t)$, at the i-th antenna may be expressed as

$$y_{i,\ell}(t) = \sum_{m=0}^{N_b-1} b[m]h_{i,\ell}(t - mT_o) + n_{i,\ell}(t), \quad i = 1, ..., N_A, \tag{2}$$

where the "channel" $h_{i,\ell}(t)$ may be expressed as $h_{i,\ell}(t) = p_{rc}(t) * h_{RF_i}(t) * f_\ell(t)$. Following the lead of [1], for the ℓ-th subband, we would like to determine a set of channel tap values $h_{i,\ell}^{(c)}[k]$ equi-spaced across the delay spread ideally satisfying

$$\begin{aligned} h_{i,\ell}(t) &= \sum_{k=1}^{Q} g_{ik}\delta_a(t - \tau_k) * \{p_{rc}(t) * f_\ell(t)\} \\ &= \sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k]\delta_a(t - k\frac{\tau_{max}}{N_h}) * \{p_{rc}(t) * f_\ell(t)\}. \end{aligned} \tag{3}$$

That is, we wish to determine $h_{i,\ell}^{(c)}[k]$, $k = 0, 1, ..., N_h$, so that the Fourier transforms of both $\sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k]\delta_a(t - k\frac{\tau_{max}}{N_h})$ and $\sum_{k=1}^{Q} g_{ik}\delta_a(t - \tau_k)\}$ over the passband of the analysis filter $f_\ell(t)$ are ideally equal. This is sufficient for our equalization purposes here.

Again, the proposed system operates at twice the symbol rate. A model for the $y_{i,\ell}[n] = y_{i,\ell}(nT_o/2)$ under noiseless conditions may be obtained by substituting $nT_o/2$ into eqn.(2) and ignoring noise (in order to develop the Cross-Relation Method).

$$\begin{aligned} y_{i,\ell}[n] &= y_{i,\ell}\left(n\frac{T_o}{2}\right) = \sum_{m=0}^{N_b-1} b[m]h_{i,\ell}(t - mT_o)\bigg|_{t=nT_o/2} \\ &= \sum_{m=0}^{N_b-1} b[m]h_{i,\ell}[n - 2m] \end{aligned} \tag{4}$$

where $h_{i,\ell}[n] = h_{i,\ell}(nT_o/2)$. Consider choosing J and N_h so that the ratio J/N_h is an integer so that, in turn,

$$\frac{\tau_{max}}{N_h} = I\frac{T_o}{2} \quad \text{where: } I = \frac{J}{N_h} = \text{integer.} \tag{5}$$

Sampling the expression for $h_{i,\ell}(t)$ (with the substitution above) at twice the symbol rate yields the following expression for the impulse response of the channel associated with the ℓ-th subband and i-th antenna:

$$\begin{aligned} h_{i,\ell}[n] &= h_{i,\ell}\left(n\frac{T_o}{2}\right) = \sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k]\delta_a(t - k\frac{\tau_{max}}{N_h}) * \{p_{rc}(t) * f_\ell(t)\}\bigg|_{t=nT_o/2} \\ &= \sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k]f_\ell(t - kI\frac{T_o}{2}) * p_{rc}(t)\bigg|_{t=nT_o/2} = \sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k]f_\ell[n - Ik] * p_{rc}[n] \\ &= \left\{\sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k]\delta[n - Ik]\right\} * \{p_{rc}[n] * f_\ell[n]\} \qquad (6) \\ &= \sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k]p_\ell[n - Ik] \qquad (7) \end{aligned}$$

where $\delta[n]$ is the Kronecker delta function and

$$p_{rc}[n] = p_{rc}\left(n\frac{T_o}{2}\right) \quad \text{and} \quad f_\ell[n] = f_\ell\left(n\frac{T_o}{2}\right) \quad \text{and} \quad p_\ell[n] = p_{rc}[n] * f_\ell[n]. \tag{8}$$

Both the model for $h_{i,\ell}[n]$ in eqn. (6) and that in eqn. (7) will be useful for later developments.

2 Blind Channel Identification

To apply the Cross-Relation (CR) method, consider decimating $y_{i,\ell}[n]$ in (4) by a factor of 2.

$$y_{i,\ell}^{(0)}[n] = y_{i,\ell}[2n] = b[n] * h_{i,\ell}^{(0)}[n], \quad \text{where} \quad h_{i,\ell}^{(0)}[n] = h_{i,\ell}[2n]. \tag{9}$$

For the sake of simplicity, assume without loss of generality that there are only $N_A = 2$ antennas. Considering $i = 1$ and $i = 2$, eqn. (9) indicates that we have the same input, the symbol sequence $b[m]$, into two different FIR filters, $h_{1,\ell}^{(0)}[n]$ and $h_{2,\ell}^{(0)}[n]$. We thus have the cross-relation:

$$y_{1,\ell}^{(0)}[n] * h_{2,\ell}^{(0)}[n] - y_{2,\ell}^{(0)}[n] * h_{1,\ell}^{(0)}[n] = 0 \tag{10}$$

Decimating the model for $h_{i,\ell}[n]$ in eqn.(7) by a factor of 2 and substituting into the equation above yields

$$\left\{ \sum_{k=0}^{N_h} h_{1,\ell}^{(c)}[k] p_\ell[2n - Ik] \right\} * y_{2,\ell}^{(0)}[n] - \left\{ \sum_{k=0}^{N_h} h_{2,\ell}^{(c)}[k] p_\ell[2n - Ik] \right\} * y_{1,\ell}^{(0)}[n] = 0. \tag{11}$$

Invoking the distributive property of convolution, the above equation may be rewritten as

$$\sum_{k=0}^{N_h} h_{1,\ell}^{(c)}[k] y_{2,\ell}^{(0,k)}[n] - \sum_{k=0}^{N_h} h_{2,\ell}^{(c)}[k] y_{1,\ell}^{(0,k)}[n] = 0 \quad \forall n, \tag{12}$$

where

$$y_{i,\ell}^{(0,k)}[n] = p_\ell[2n - Ik] * y_{i,\ell}^{(0)}[n], \quad k = 0, 1, ..., N_h, \quad i = 1, 2. \tag{13}$$

The relation in eqn. (12) may be expressed in matrix form as

$$\left[\mathbf{Y}_{2,\ell}^{(0)} \vdots - \mathbf{Y}_{1,\ell}^{(0)} \right] \mathbf{h}_\ell^{(c)} = \mathbf{0} \qquad 1 \le \ell \le M, \tag{14}$$

where

$$\mathbf{h}_\ell^{(c)} = \left[h_{1,\ell}^{(c)}[0], ..., h_{1,\ell}^{(c)}[N_h], h_{2,\ell}^{(c)}[0], ..., h_{2,\ell}^{(c)}[N_h] \right]^T \tag{15}$$

and

$$\mathbf{Y}_{i,\ell}^{(0)} = \begin{bmatrix} y_{i,\ell}^{(0,0)}[0] & y_{i,\ell}^{(0,1)}[0] & \cdots & y_{i,\ell}^{(0,N_h)}[0] \\ y_{i,\ell}^{(0,0)}[1] & y_{i,\ell}^{(0,1)}[1] & \cdots & y_{i,\ell}^{(0,N_h)}[1] \\ \vdots & \vdots & & \vdots \\ y_{i,\ell}^{(0,0)}[N_s] & y_{i,\ell}^{(0,1)}[N_s] & \cdots & y_{i,\ell}^{(0,N_h)}[N_s] \end{bmatrix} \tag{16}$$

$i = 1, 2$, where $N_s + 1 \le N_b$ is the number of symbols used to estimate the channels.

Similar to the development in[1], the performance may be improved by generating additional equations by employing the other polyphase component of the ℓ-th subband output of the two antennas.

$$y_{i,\ell}^{(1)}[n] = y_{i,\ell}[2n + 1] = b[n] * h_{i,\ell}^{(1)}[n], \quad \text{where} \quad h_{i,\ell}^{(1)}[n] = h_{i,\ell}[2n + 1]. \tag{17}$$

Similar to the previous development, we have the cross-relation:

$$y_{1,\ell}^{(1)}[n] * h_{2,\ell}^{(1)}[n] - y_{2,\ell}^{(1)}[n] * h_{1,\ell}^{(1)}[n] = 0 \tag{18}$$

The critical observation for improved performance is that the same coefficients $h_{i,\ell}^{(c)}[k]$ characterize both polyphase components of $h_{i,\ell}[n]$. Decimating the model for $h_{i,\ell}[n]$ in eqn.(7) by a factor of 2 starting at $n = 1$ and substituting into the equation above yields

$$\left\{ \sum_{k=0}^{N_h} h_{1,\ell}^{(c)}[k] p_\ell[2n + 1 - Ik] \right\} * y_{2,\ell}^{(1)}[n] - \left\{ \sum_{k=0}^{N_h} h_{2,\ell}^{(c)}[k] p_\ell[2n + 1 - Ik] \right\} * y_{1,\ell}^{(1)}[n] = 0 \tag{19}$$

which may ultimately be expressed as

$$\sum_{k=0}^{N_h} h_{1,\ell}^{(c)}[k] y_{2,\ell}^{(1,k)}[n] - \sum_{k=0}^{N_h} h_{2,\ell}^{(c)}[k] y_{1,\ell}^{(1,k)}[n] = 0 \quad \forall \text{n}, \tag{20}$$

where

$$y_{i,\ell}^{(1,k)}[n] = p_\ell[2n+1-Ik] * y_{i,\ell}^{(1)}[n], \quad k = 0, 1, \ldots, N_h, \quad i = 1, 2. \tag{21}$$

The relation in eqn. (20) may be expressed in matrix form as

$$\left[\mathbf{Y}_{2,\ell}^{(1)} \vdots - \mathbf{Y}_{1,\ell}^{(1)}\right] \mathbf{h}_\ell^{(c)} = \mathbf{0} \quad 1 \le \ell \le M, \tag{22}$$

where

$$\mathbf{Y}_{i,\ell}^{(1)} = \begin{bmatrix} y_{i,\ell}^{(1,0)}[0] & y_{i,\ell}^{(1,1)}[0] & \cdots & y_{i,\ell}^{(1,N_h)}[0] \\ y_{i,\ell}^{(1,0)}[1] & y_{i,\ell}^{(1,1)}[1] & \cdots & y_{i,\ell}^{(1,N_h)}[1] \\ \vdots & \vdots & & \vdots \\ y_{i,\ell}^{(1,0)}[N_s] & y_{i,\ell}^{(1,1)}[N_s] & \cdots & y_{i,\ell}^{(1,N_h)}[N_s] \end{bmatrix} \tag{23}$$

Appending eqn.(22) to eqn.(14) yields

$$\underbrace{\begin{bmatrix} \mathbf{Y}_{2,\ell}^{(0)} & -\mathbf{Y}_{1,\ell}^{(0)} \\ \mathbf{Y}_{2,\ell}^{(1)} & -\mathbf{Y}_{1,\ell}^{(1)} \end{bmatrix}}_{\mathbf{X}_\ell(N_s)} \mathbf{h}_\ell^{(c)} = \mathbf{0} \quad 1 \le \ell \le M \tag{24}$$

In the practical case where the channels are corrupted by noise, we follow the lead of Xu et al [3] and choose $\mathbf{h}_\ell^{(c)}$ as the least significant right singular vector of $\mathbf{X}_\ell(N_s)$.

3 Maximization of In-Band Energy

Recall the cross-relation in (10) and that the output of each analysis filter $f_\ell[n]$ is maintained at twice the symbol rate. For the narrowband case, we observed in [1] that oversampling leads to improved performance but also trivial solutions as well in which the cross-relation is satisfied by making both $y_{1,\ell}^{(0)}[n] * h_{2,\ell}^{(0)}[n]$ and $y_{2,\ell}^{(0)}[n] * h_{1,\ell}^{(0)}[n]$ individually zero for all n (approximately zero in practice.) This occurs when the matrix $\mathbf{X}_\ell(N_s)$ in eqn. (24) has a number of small singular values and the right singular vector associated with the smallest singular value yields $h_{1,\ell}^{(0)}[n]$ and $h_{2,\ell}^{(0)}[n]$ that are band-reject filters with respective DTFT's approximately equal to zero for $\omega_{1\ell} \le \omega \le \omega_{2\ell}$, where $(\omega_{1\ell}, \omega_{2\ell})$ is the passband of the analysis filter $f_\ell[n]$.

We propose a maximum in-band energy post-processing step to avert this phenomenon. Let $\mathbf{V}_\ell$ be a matrix containing the right singular vectors of $\mathbf{X}_\ell(N_s)$ associated with the K smallest singular values. The vector $\mathbf{h}_\ell^{(c)} = \mathbf{V}_\ell \boldsymbol{\beta}_\ell$ basically satisfies the cross-relation for any value of the $K \times 1$ vector $\boldsymbol{\beta}_\ell$. The proper linear combination $\boldsymbol{\beta}_\ell$ is found by maximizing the in-band energy of $\sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k]\delta[n-Ik]$ via the solution to

$$\max_{\mathbf{h}_\ell^{(c)}} \frac{\displaystyle\sum_i \begin{array}{c}\text{in-band energy for channel } i \\ \text{corresponding to subband filter } f_\ell[n]\end{array}}{\displaystyle\sum_i \text{total energy for channel } i} = \max_{\mathbf{h}_\ell^{(c)}} \frac{\displaystyle\sum_{i=1}^{N_A} \int_{\omega_{1\ell}}^{\omega_{2\ell}} \left| \sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k] e^{-jI\omega k} \right|^2 d\omega}{\displaystyle\sum_{i=1}^{N_A} \int_{-\pi}^{\pi} \left| \sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k] e^{-jI\omega k} \right|^2 d\omega} \tag{25}$$

After some manipulation, eqn.(25) may be formulated by determining $\boldsymbol{\beta}_\ell$ as the solution to

$$\max_{\boldsymbol{\beta}_\ell} \frac{\boldsymbol{\beta}_\ell{}^H \mathbf{V}_\ell{}^H \mathbf{B}_\ell{}^{(2)} \mathbf{V}_\ell \boldsymbol{\beta}_\ell}{\boldsymbol{\beta}_\ell{}^H \mathbf{V}_\ell{}^H \mathbf{V}_\ell \boldsymbol{\beta}_\ell} \tag{26}$$

where for $N_A = 2$ antennas

$$\mathbf{B}_\ell{}^{(2)} = \begin{bmatrix} \mathbf{F}_\ell & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_\ell \end{bmatrix} \quad \text{and} \quad \mathbf{F}_\ell = \int_{\omega_{1\ell}}^{\omega_{2\ell}} \mathbf{f}(\omega)\mathbf{f}^H(\omega)d\omega \tag{27}$$

with $\mathbf{f}(\omega) = [1, e^{-jI\omega}, ..., e^{-jIN_h\omega}]^T$. A closed-form solution to eqn. (25) may be computed by setting $\boldsymbol{\beta}_\ell$ in $\mathbf{h}_\ell^{(c)} = \mathbf{V}_\ell\boldsymbol{\beta}_\ell$ equal to the most significant right singular vector of $\mathbf{V}_\ell^H\mathbf{B}_\ell^{(2)}\mathbf{V}_\ell$.

4 Subband Zero-Forcing Equalizers

Again, we have assumed two distinct antennas exist at the receiver. The baseband output of each antenna has been decomposed into M subbands. We have applied a variation of the CR method in each subband to determine the respective frequency response of the multipath propagation channel for each of the two antennas across the given subband. The next step is to effect a zero forcing equalizer in each subband across the two antennas.

Given $\mathbf{h}_{i,\ell}^{(c)}=[h_{i,\ell}^{(c)}[0], ..., h_{i,\ell}^{(c)}[N_h]]^T$, $i = 1, 2$, the channel impulse response for the i-th antenna associated with two times per symbol sampling is formed as $\sum_{k=0}^{N_h} h_{i,\ell}^{(c)}[k]p_\ell[n - Ik]$, $i = 1, 2$. Zero-forcing equalization may be effected at a lower rate by constraining the equalizing filters $g_{i,\ell}[n]$ satisfying the zero forcing constraint

$$h_{1,\ell}[n] * g_{1,\ell}[n] + h_{2,\ell}[n] * g_{2,\ell}[n] = p_\ell[n - D] \tag{28}$$

to be of the form $\sum_{k=0}^{N_h} g_{i,\ell}^{(c)}[k]\delta[n - Ik]$, which describes a sequence having $I - 1$ zeroes between every two successive values of $g_{i,\ell}^{(c)}[k]$. Since the DTFT of the left hand side of (28) need only have flat magnitude and linear phase over the passband of $f_\ell[n]$, the criterion for designing the equalizing filters $g_{i,\ell}[n]$ is

$$\min_{\mathbf{g}_{1,\ell}^{(c)}, \mathbf{g}_{2,\ell}^{(c)}} \int_{\omega_{1\ell}}^{\omega_{2\ell}} F_\ell(\omega) \left| \sum_{i=1}^{2} H_{i,\ell}^{(c)}(\omega)G_{i,\ell}^{(c)}(\omega) - e^{-j\omega D_e} \right|^2 d\omega \tag{29}$$

where $\mathbf{g}_{i,\ell}^{(c)} = [g_{i,\ell}^{(c)}[0], ..., g_{i,\ell}^{(c)}[N_g]]^T$ and $F_\ell(\omega)$ is the discrete-time fourier transform of $f_\ell[n]$. Further,

$$H_{i,\ell}^{(c)}(\omega) = \mathbf{f}_h^H(\omega)\mathbf{h}_{i,\ell}^{(c)} \quad \text{and} \quad G_{i,\ell}^{(c)}(\omega) = \mathbf{f}_g^H(\omega)\mathbf{g}_{i,\ell}^{(c)} \tag{30}$$

where

$$\mathbf{f}_h(\omega) = [1, e^{jI\omega}, \cdots, e^{jN_hI\omega}]^T, \quad \mathbf{f}_g(\omega) = [1, e^{jI\omega}, \cdots, e^{jN_gI\omega}]^T \tag{31}$$

Thus, eqn.(29) may be expressed in a matrix form as

$$\min_{\mathbf{g}_\ell^{(c)}} \ \mathbf{g}_\ell^{(c)^H} \mathbf{F}_{hh}^{(\ell)} \mathbf{g}_\ell^{(c)} - \mathbf{f}_{hh}^{(\ell)^H} \mathbf{g}_\ell^{(c)} - \mathbf{g}_\ell^{(c)^H} \mathbf{f}_{hh}^{(\ell)} + \ \text{RC} \tag{32}$$

where

$$\mathbf{g}_\ell^{(c)} = [\mathbf{g}_{1,\ell}^{(c)^T} \ \ \mathbf{g}_{2,\ell}^{(c)^T}]^T \quad \text{and} \quad \text{RC} = \int_{\omega_{1\ell}}^{\omega_{2\ell}} F_\ell(\omega) d\omega \tag{33}$$

$$\mathbf{F}_{hh}^{(\ell)} = \begin{bmatrix} \mathbf{F}_{h_1 h_1}^{(\ell)} & \mathbf{F}_{h_1 h_2}^{(\ell)} \\ \mathbf{F}_{h_2 h_1}^{(\ell)} & \mathbf{F}_{h_2 h_2}^{(\ell)} \end{bmatrix} \quad \text{and} \quad \mathbf{f}_{hh}^{(\ell)} = [\mathbf{f}_{h_1}^{(\ell)^T} \ \mathbf{f}_{h_2}^{(\ell)^T}]^T \tag{34}$$

where

$$\mathbf{F}_{h_i h_j}^{(\ell)} = \int_{\omega_{1\ell}}^{\omega_{2\ell}} F_\ell(\omega) \mathbf{f}_g(\omega) \mathbf{h}_{i,\ell}^{(c)^H} \mathbf{f}_h(\omega) \mathbf{f}_h^H(\omega) \mathbf{h}_{j,\ell}^{(c)} \mathbf{f}_g^H(\omega) d\omega \tag{35}$$

$$\mathbf{f}_{h_i}^{(\ell)} = \int_{\omega_{1\ell}}^{\omega_{2\ell}} F_\ell(\omega) \mathbf{f}_g(\omega) \mathbf{h}_{i,\ell}^{(c)^H} \mathbf{f}_h(\omega) e^{-j\omega D} d\omega \tag{36}$$

A simple closed-form solution exists to the problem in eqn.(29) is $\mathbf{g}_\ell^{(c)} = \mathbf{F}_{hh}^{-1} \mathbf{f}_{hh}^{(\ell)}$. The derivation is not included here due to space limitations.

5 Coherent Phasing of Subband Channel Estimates

The equalized signal for the ℓ-th subband is synthesized as

$$y_\ell^{eq}[n] = y_{1,\ell}[n] * g_{1,\ell}[n] + y_{2,\ell}[n] * g_{2,\ell}[n]. \tag{37}$$

Since this (2-channel) zero-forcing equalizer operates at twice the symbol rate, in the case of perfect subband channel matching and no noise, $y_\ell^{eq}[n]$ is ideally of the form

$$\begin{aligned} y_\ell^{eq}[n] &= \alpha_\ell \sum_{m=0}^{N_b-1} b[m] p_\ell[n-2m] && (38) \\ &= \sum_{m=0}^{N_b-1} b[m] p_{rc}[n-2m] * \{\alpha_\ell f_\ell[n]\}, && (39) \end{aligned}$$

where α_ℓ is an unknown complex scalar that arises from the "blind" nature of the algorithm: $\mathbf{h}_\ell^{(c)}$ is computed as the right singular of $\mathbf{X}_\ell(N_s)$ which is only unique to within a complex-valued scalar. A complicating factor is that α_ℓ is generally different for each subband.

Consider the following interpretation of eqn. (39): $y_\ell^{eq}[n]$ and $y_{\ell+1}^{eq}[n]$ are obtained by passing the same input, $\sum_{m=0}^{N_b-1} b[m] p_{rc}[n-2m]$, through two different FIR filters with respective impulse responses $\alpha_\ell f_\ell[n]$ and $\alpha_{\ell+1} f_{\ell+1}[n]$. This yields the following cross-relation:

$$\alpha_\ell \, y_{\ell+1}^{eq}[n] * f_\ell[n] - \alpha_{\ell+1} \, y_\ell^{eq}[n] * f_{\ell+1}[n] = 0 \tag{40}$$

This cross-relation is exploited to determine the 2×1 vector $\boldsymbol{\alpha}_{\ell,\ell+1} = [\alpha_\ell \ , \ \alpha_{\ell+1}]^T$ as follows. First, perform the following convolutions:

$$z_{\ell,\ell+1}[n] = y_\ell^{eq}[n] * f_{\ell+1}[n] \quad \text{and} \quad z_{\ell+1,\ell}[n] = y_{\ell+1}^{eq}[n] * f_\ell[n]. \tag{41}$$

Note that it is important for the subbands to have substantial overlap or the result of each of the above two convolutions will be negligible. Defining $\mathbf{z}_{\ell,\ell+1} = [z_{\ell,\ell+1}[0], ..., z_{\ell,\ell+1}[N_\alpha]]^T$ and $\mathbf{z}_{\ell+1,\ell} = [z_{\ell+1,\ell}[0], ..., z_{\ell+1,\ell}[N_\alpha]]^T$, the cross-relation for N_α successive values of n may be cumulatively expressed as

$$\mathbf{Z}_{\ell,\ell+1} \boldsymbol{\alpha}_{\ell,\ell+1} = \mathbf{0}, \quad \text{where:} \ \mathbf{Z}_{\ell,\ell+1} = [\mathbf{z}_{\ell+1,\ell} \vdots - \mathbf{z}_{\ell,\ell+1}] \tag{42}$$

$\alpha_{\ell,\ell+1}$ is thus computed as the the right singular vector of $\mathbf{Z}_{\ell,\ell+1}$ associated with the smallest singular value. Eqn.(42) can be efficiently solved by pre-multiplying both sides by $\mathbf{Z}_{\ell,\ell+1}^H$. Cumulatively, all α_ℓ, $1 \leq \ell \leq M$, may be solved for simultaneously by forming the following $(2M-2) \times M$ matrix.

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_{2,1}^H\mathbf{z}_{2,1} & \mathbf{z}_{2,1}^H\mathbf{z}_{1,2} & 0 & \cdots & 0 \\ \mathbf{z}_{1,2}^H\mathbf{z}_{2,1} & \mathbf{z}_{1,2}^H\mathbf{z}_{1,2} & 0 & \cdots & 0 \\ 0 & \mathbf{z}_{3,2}^H\mathbf{z}_{3,2} & \mathbf{z}_{3,2}^H\mathbf{z}_{2,3} & 0 & \cdots \\ 0 & \mathbf{z}_{2,3}^H\mathbf{z}_{3,2} & \mathbf{z}_{2,3}^H\mathbf{z}_{2,3} & 0 & \cdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{z}_{M,M-1}^H\mathbf{z}_{M,M-1} & \mathbf{z}_{M,M-1}^H\mathbf{z}_{M-1,M} \\ 0 & \cdots & 0 & \mathbf{z}_{M-1,M}^H\mathbf{z}_{M,M-1} & \mathbf{z}_{M-1,M}^H\mathbf{z}_{M-1,M} \end{bmatrix} \tag{43}$$

$\boldsymbol{\alpha}_\ell = [\alpha_1, ..., \alpha_M]^T$ is computed as the eigenvector of $\mathbf{Z}$ associated with the smallest eigenvalue.

The reconstructed signal is formed as in Figure 1 as $\hat{y}[n] = \sum_{\ell=1}^{M} \frac{1}{\alpha_\ell} y_\ell^{eq} * e_\ell[n]$, where $e_\ell[n]$, $\ell = 1, ..., M$, represent a bank of synthesis filters designed so that

$$DTFT\left\{\sum_{\ell=1}^{M} f_\ell[n] * e_\ell[n]\right\} = 1, \text{ for } |\omega| < \frac{1+\beta}{2}\pi \tag{44}$$

The reconstructed signal $\hat{y}[n]$ is an estimate of $y[n] = \sum_{m=0}^{N_b-1} b[m]p_{rc}[n-D-2m]$.

6 Simulations

In the simulations, the multipath signals arrived at -5^o, 0^o, and 5^o with complex amplitudes of $0.8e^{2\pi/3}$, 1, $0.9e^{\pi/4}$, respectively, and time delays of 10, 0, and 6 micro-seconds (μs). The array was composed of N_A=2 antennas separated by three wavelengths. The symbol alphabet was QPSK. The symbol period was 1 μ s; the symbol rate was 1 MHz. The symbol waveform had a square-root raised cosine spectrum with β=0.35. The RF bandwidth was 1.35 MHz.

Each analysis filter, $f_\ell[n]$, $\ell = 1, ..., 15$, had a raised cosine spectrum with a different center frequency and passed roughly a 200 KHz subband as plotted in Figures 2(a) and 2(b). This choice of analysis filters averted the need for synthesis filters, $e_\ell[n]$, since the sum of the spectra in either Figure 2(a) or 2(b) is spectrally flat over the 1.3 MHz baseband bandwidth of the desired signal.

The subbanding based blind multi-channel identification algorithm presented herein was applied to a block of N_s = 200 symbols at an SNR of 25 dB. Figure 3(a), 3(b) and 3(c) plot the magnitude of the DTFT below for three different subbands

$$\hat{p}_\ell[n] = h_{1,\ell}[n] * \hat{g}_{1,\ell}[n] + h_{2,\ell}[n] * \hat{g}_{2,\ell}[n], \text{for } \ell = -1, 0, 1$$

using the actual values of $h_{1,\ell}[n]$ and $h_{2,\ell}[n]$ and the estimated values of $g_{1,\ell}[n]$ and $g_{2,\ell}[n]$ obtained for a typical trial run. For these subbands, the DTFT of the ideal value of $h_{1,\ell}[n] * g_{1,\ell}[n] + h_{2,\ell}[n] * g_{2,\ell}[n]$, $p_\ell[n] = p_{rc}[n] * f_\ell[n]$, is also plotted in each of the figures for comparison purposes; the match is pretty close in all three cases. The estimated results $\hat{p}_\ell[n]$, $\ell = -7, ..., 7$, were combined via the estimated coefficients, α_ℓ, as $\hat{p}_{rc}[n] = \sum_{\ell=-7}^{7} \frac{1}{\hat{\alpha}_\ell} \hat{p}_\ell[n] * e_\ell[n]$ to yield the result is plotted in Fig. 3(d). Again, given our judicious choice of analysis filters, the synthesis filters, $e_\ell[n]$, required were simply Kronecker delta functions. The resulting reconstructed raised cosine spectrum is observed to have moderate distortions relative to the ideal raised cosine spectrum. The 200 reconstructed symbols are plotted in Fig. 4(b) and are observed to highly cluster about four points commensurate with QPSK. This is in contrast to the received signal constellation at the first antenna plotted in Fig. 4(a).

References

[1] M. D. Zoltowski, D. Tseng, and T. A. Thomas, "On the Use of Basis Functions in Blind Equalization Based on Deterministic Least Squares," *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, pp. 816-822, Nov. 1997.

[2] M. D. Zoltowski, D. Tseng, "A Weighted Energy Concentration Criterion for Improving the Performance of Deterministic Least Squares Blind Channel Identification," to appear *Proc. Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1998.

[3] G. Xu, H. Liu, L. Tong & T. Kailath, "A Least-Squares Approach to Blind Channel Identification," *IEEE Trans. on Signal Processing,* pp. 2982-2993, Dec. 1995.

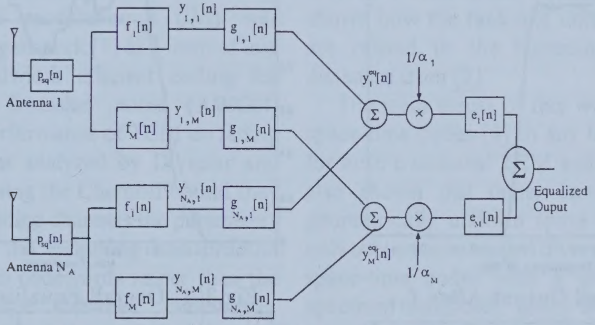

$f_1[n]$
$y_{1,1}[n]$
$g_{1,1}[n]$
$P_{sq}[n]$
Antenna 1
$f_M[n]$
$y_{1,M}[n]$
$g_{1,M}[n]$
$f_1[n]$
$y_{N_A,1}[n]$
$g_{N_A,1}[n]$
$P_{sq}[n]$
Antenna N_A
$f_M[n]$
$y_{N_A,M}[n]$
$g_{N_A,M}[n]$
$1/\alpha_1$
$y_1^{eq}[n]$
Σ
×
$e_1[n]$
Σ
Equalized Ouput
Σ
×
$e_M[n]$
$y_M^{eq}[n]$
$1/\alpha_M$

Fig. 1. Block Diagram of Space-Time Equalizer.

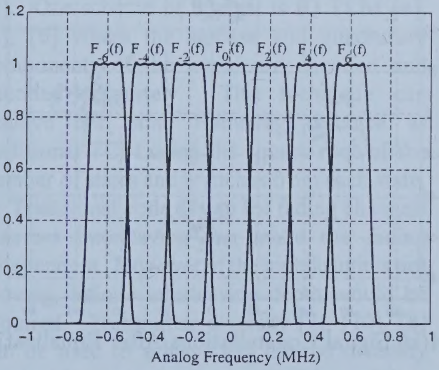


Fig.2(a) Filterbank 1

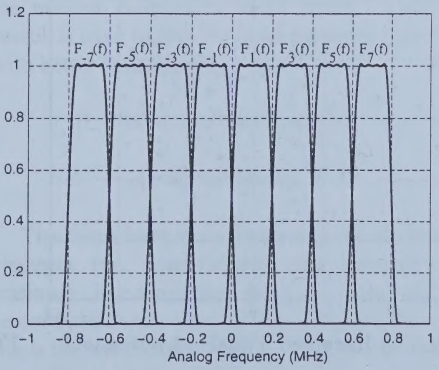


Fig.2(b) Filterbank 2

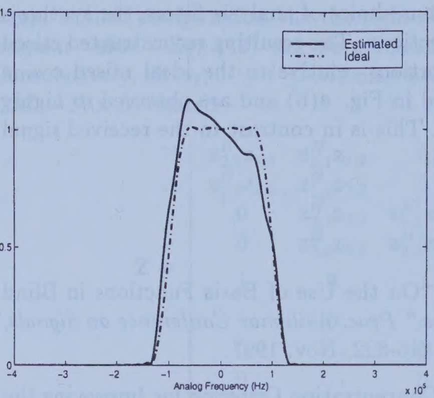


Fig.3(a) Equalized Output After f_0

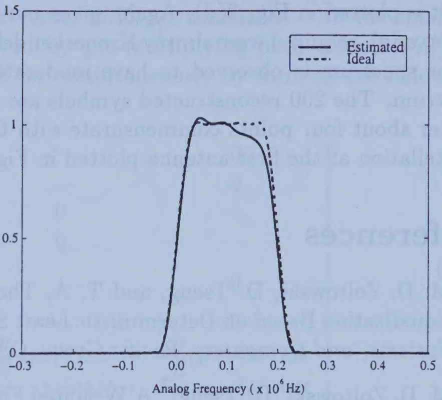


Fig.3(b) Equalized Output After f_1

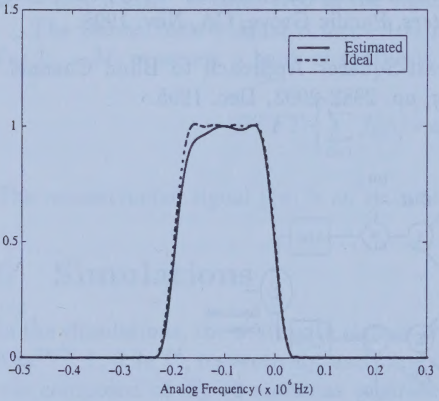


Fig.3(c) Equalized Output After f_{-1}

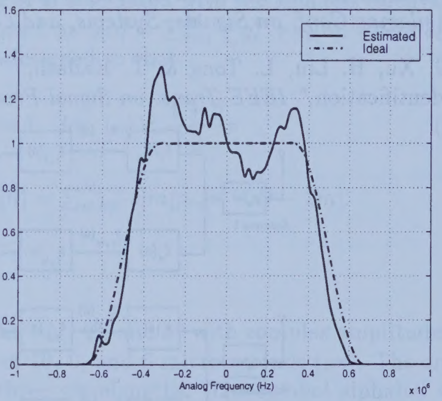


Fig.3(d) Overall Equalized Output

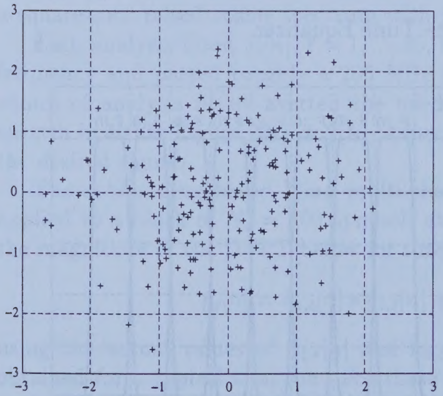


Fig.4(a) Received Signals at antenna 1

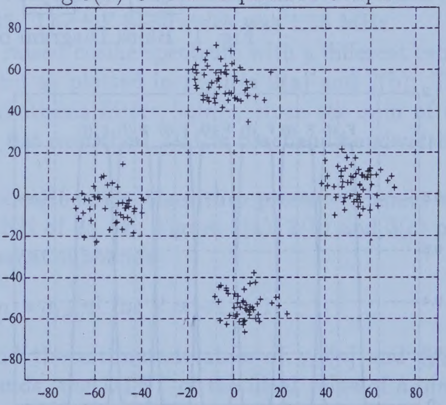


Fig.4(b) Signal Constellations After Equalization

Further Results on Space-Time Coding for Rayleigh Fading

J. Grimm (jimm@ecn.purdue.edu), M. P. Fitz (fitz.7@osu.edu)
and J. V. Krogmeier (jvk@ecn.purdue.edu)
School of Electrical Engineering, Purdue University, W. Lafayette, IN
and Department of Electrical Engineering, Ohio State University, Columbus, OH

Abstract – **This paper presents further results in the geometric uniformity of space-time codes. Extensions of previously proposed AT&T design rules for full diversity space-time codes are discussed that can be applied to any level of diversity. The results of a search for good space-time codes are presented.**

I. INTRODUCTION

Trellis coded modulation (TCM) was introduced by Ungerboeck [1] as a form of low complexity, bandwidth efficient coding for additive white Gaussian noise (AWGN) channels. The performance of TCM on fading channels was first analyzed by Divsalar and Simon [2], and using the Chernoff bound they proved that for fading channels the parameters to optimize are the Hamming and product distances between codewords rather than the Euclidean distance between codewords. Numerous researchers have designed many trellis codes for independently fading channels [3]. One scheme of interest is IQ TCM [4], [5], [6] where the inphase and quadrature components of the signal are encoded and decoded separately. This technique can achieve the same Hamming distance as traditional TCM using the square root of the number of states and branches from each state.

Traditional code design for fading channels assumes interleaving can make the fading independent. However, if the mobile unit is not moving, infinite interleaving depth would be required. Transmitter diversity [7], [10], [8] can be used to achieve any desired diversity level without interleaving, even if the mobile is not moving. A powerful TCM scheme for transmitter diversity called space-time codes was proposed in [8] for two levels of diversity.

Guey et al. [9], [10] derived the exact pairwise error performance for transmitter diversity, and were the first to prove that the diversity level is equal to the rank of the codeword difference matrices. Their derivation also showed the relationship between performance and the singular values of the codeword difference matrix. The authors in [8] derived a determinant criterion. In [11] it is shown how the rank and determinant criteria are related to the Hamming and product distances from [2].

The main thrust of this work is to extend space-time codes [8] to any level of diversity for both traditional TCM and IQ TCM. It is also shown that under certain restrictions, geometrically uniform space-time codes can only achieve one level of diversity, so designing space-time codes with a uniform distance spectrum is difficult. Delay diversity [12], [8] is generalized to any throughput and diversity level, and it is proven that delay diversity uses the minimal number of trellis states. Code search is used to find higher complexity codes with better distance properties.

II. PROBLEM FORMULATION

A. Signal Model

The focus here is on frequency flat fading channels (i.e., narrowband data communications). For simplicity, the code design will be confined to non-time varying channels, since this is the worst case scenario. This means that the fading is 100% correlated in time, so if the mobile is stopped in a deep fade interleaving won't help. If the mobile unit is moving and time diversity is available, the codes designed here will have at least as good performance as in the non-time varying case, and most likely better performance on average. The design considers antenna (spatial) diversity and time diversity, hence the name space-time code. The system has L_t transmitter antennas and L_r receiver antennas, and it is assumed that the fading across the antennas is uncorrelated. Since the receiver antennas are uncorrelated, the total diversity level achieved will be L_r times the diversity of the rest of the system. Consequently the code design can focus on the L_t transmitter antennas and N_c time instants. When a length N_c codeword is transmitted, the received signal $x_\alpha(n)$ may be modeled as

$$x_\alpha(n) = \sqrt{E_b}\sum_{l=1}^{L_t} d_\alpha(n,l)c(l) + n(n), \quad n=1{:}N_c \quad \mathbf{(1)}$$

where E_b is the total energy per information bit, the transmitted symbols at time n and on antenna l, $d_\alpha(n,l)$, are normalized to unit average energy per information bit, $c(l)$ is the unit variance complex Gaussian channel distortion from the l^{th} transmitter antenna, and the complex AWGN $n(n)$ has variance N_0. Equation (1) can be written more compactly in matrix notation:

$$\vec{x}_\alpha = \sqrt{E_b}D_\alpha\vec{c} + \vec{n} \quad \mathbf{(2)}$$

where $\vec{x}_\alpha$ and $\vec{n}$ are $N_c\times 1$ vectors, D_α is an $N_c\times L_t$ matrix with elements $d_\alpha(n,l)$, and $\vec{c}$ is $L_t\times 1$. Throughout this paper it is assumed that $N_c\geq L_t$.

To summarize from [8], [11] the two most important design criteria for transmitter diversity systems are the minimum 1) effective Hamming distance (e_H) and 2) effective product distance (e_P) among all codeword differences $D_{\alpha\beta}=D_\alpha-D_\beta$. For the simple model presented here these are defined as

$$\begin{aligned} e_P(D_{\alpha\beta}) &\triangleq \prod_{n\in\eta}\lambda_n(C_1) = \prod_{n\in\eta}\lambda_n(C_2) \\ e_H(D_{\alpha\beta}) &\triangleq \operatorname{rank}(C_1) = \operatorname{rank}(C_2) = \operatorname{rank}(D_{\alpha\beta}) \end{aligned} \quad \mathbf{(3)}$$

where $C_1=D_{\alpha\beta}D_{\alpha\beta}^H$, $C_2=D_{\alpha\beta}^HD_{\alpha\beta}$, $\lambda_n(A)$ is the n^{th} nonzero eigenvalue of the matrix A, and η is the set of all nonzero eigenvalues. The terms e_H and rank will be used interchangeably in the sequel, and we define a full rank codeword difference matrix to have $e_H=L_t$. Note that when $D_{\alpha\beta}$ is full rank, $e_P(D_{\alpha\beta}) = \det(C_2)$.

The effective Hamming distance is the diversity level achieved (slope of the error probability versus signal to noise ratio (SNR) at high SNR), and the effective product distance gives vertical displacement to the error probability curve. In [8], e_H is called the rank criterion and e_P is called the determinant criterion, but the terminology used here emphasizes the relation to the original metrics in [2] for single antenna fading channel code design. The metrics e_H and e_P operate on the eigenvalues of C_1 in the same way the Hamming and product distances [2] operate on codeword differences directly in ideal interleaved single antenna TCM. It should be noted that the product distance does not meet the formal requirements of a norm (the triangle inequality does not hold), but the term "distance" will still be used to maintain consistency with the well developed literature in the area of coding for fading channels. The worst case e_P out of all codeword differences of a code is denoted $e_{P\min}$. Likewise, the minimum e_H among all codeword pairs is denoted $e_{H\min}$. The goal of code design in this paper is to find codes with the best $e_{P\min}$ out of all codes with $e_{H\min}=L_t$.

B. Terminology

A TCM encoder for space-time coding is shown in Fig. 1. At each symbol time n, k information bits are fed into the convolutional encoder. The encoder has ν binary memory elements spread across the k input bits, hence the code has $V=2^\nu$ states. Every symbol time the encoder outputs an $L_t \times 1$ vector $\vec{J}(n,:)$, whose components are mapped to a signal constellation (see Fig. 2) with $M=2^m$ possible symbols. The output symbol mapping is denoted $\vec{d}(n,:) = h(\vec{J}(n,:))$. At each time, the output vector $\vec{d}(n,:)=[d(n,1)\ d(n,2)\cdots d(n,L_t)]$ has L_t symbols, one for each of the L_t antennas. When IQ TCM is used, there are m bits in the I channel and m bits in Q, so the constellation has M^2 symbols.

Notice that all input bits are coded, because no diversity can be achieved with parallel transitions unless multiple TCM (MTCM) [13] is employed, and MTCM codes are not considered here. When MTCM is not used, the maximum attainable diversity level L_D is bounded by $L_D \le (\lfloor v/k \rfloor + 1)L_r$ [14], so full diversity codes require $\lfloor v/k \rfloor \ge L_t - 1$.

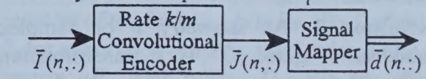

Rate k/m
Convolutional
Encoder
Signal
Mapper
$\bar{I}(n,:)$
$\bar{J}(n,:)$
$\bar{d}(n,:)$

Fig. 1. TCM encoder block diagram.

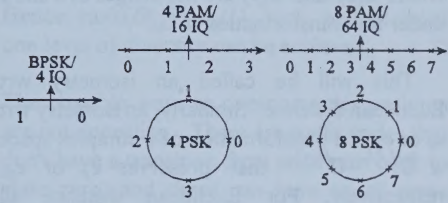

BPSK/
4 IQ
1
0
4 PAM/
16 IQ
0
1
2
3
8 PAM/
64 IQ
0
1
2
3
4
5
6
7
4 PSK
1
2
0
3
8 PSK
2
3
1
4
0
5
7
6

Fig. 2. Signal constellations used in this work. The PAM constellations can be used either for PAM or IQ-QAM.

A trellis for a typical space-time code is shown in Fig. 3. This is the L_t=2, k=2, v=3, m=2 TCM from [8]. Each branch has L_t symbols, one for each transmitter antenna. From each state, the uppermost branch corresponds to all input bits equal to zero, and the lower branches correspond to stepping through the binary numeric representation of the input bits.

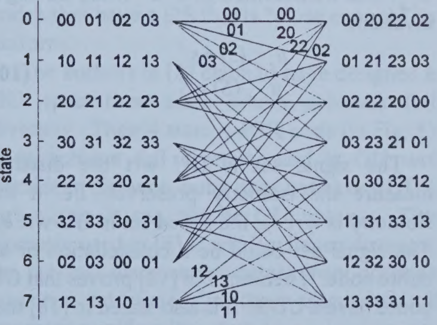

state
0
1
2
3
4
5
6
7
00 01 02 03
10 11 12 13
20 21 22 23
30 31 32 33
22 23 20 21
32 33 30 31
02 03 00 01
12 13 10 11
00
01
02
03
00
20
22
02
12
13
10
11
00 20 22 02
01 21 23 03
02 22 20 00
03 23 21 01
10 30 32 12
11 31 33 13
12 32 30 10
13 33 31 11

Fig. 3. AT&T's 8 state, 4-PSK, L_t=2 TCM.

Since space-time codes have a similar form to traditional single antenna codes the same notation will be used. Specifically in the sequel the idea of the all-zeros path, an error event path (EEP), codeword differences, and distance spectrum (DS) will be important and have identical meaning as when used in single antenna TCM. For the purpose of code analysis, it is desirable to use codes where the DS is the same regardless of the transmitted codeword. This will be called a uniform distance spectrum (UDS). A catastrophic code has EEPs where a finite number of channel symbol differences are produced by an infinite number of bit differences. Necessary and sufficient conditions for obtaining noncatastrophic feedforward convolutional (FFC) codes with k inputs and L_t outputs are given in [15]. These same conditions hold for the codes in this work.

C. Space-Time TCM

The structure of FFC TCM codes may be generalized to support space-time codes as shown in Fig. 4. Such an encoder is required for each of the binary input bits $I_i(n)$, i=1:k. The encoder memory for each input may be written $\bar{I}_{v_i}(n,i) = [I_i(n)\ I_i(n-1)\ \cdots\ I_i(n-v_i)]^t$. The tap weights $g_i(l,\tau)$ and the output symbol labels $J_i(n,l)$ are elements of $GF(2^m)$. The encoder for each input bit may have different memory length v_i, and the total memory of the encoder is $v = \sum_{i=1}^{k} v_i$. To keep the coding gain as uniform as possible for each bit, we only consider $\lfloor v/k \rfloor \le v_i \le \lceil v/k \rceil$.

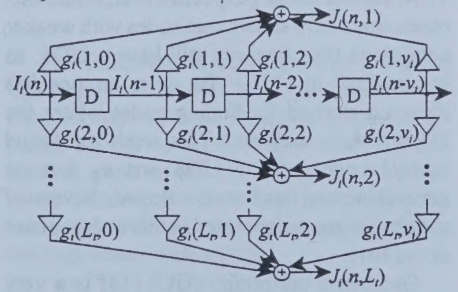

$J_i(n,1)$
$g_i(1,0)$
$g_i(1,1)$
$g_i(1,2)$
$g_i(1,v_i)$
$I_i(n)$
D
$I_i(n-1)$
D
$I_i(n-2)$
D
$I_i(n-v_i)$
$g_i(2,0)$
$g_i(2,1)$
$g_i(2,2)$
$g_i(2,v_i)$
$J_i(n,2)$
$g_i(L_r,0)$
$g_i(L_r,1)$
$g_i(L_r,2)$
$g_i(L_r,v_i)$
$J_i(n,L_t)$

Fig. 4. Encoder for i^{th} input bit.

The FFC encoder may be represented by k generator matrices

$$G_i=\begin{bmatrix} g_i(1,0) & g_i(1,1) & \cdots & g_i(1,v_i) \\ g_i(2,0) & g_i(2,1) & \cdots & g_i(2,v_i) \\ \vdots & \vdots & \ddots & \vdots \\ g_i(L_t,0) & g_i(L_t,1) & \cdots & g_i(L_t,v_i) \end{bmatrix},\ i=1{:}k \quad (4)$$

The coded bit vector $\vec{J}(n,:)$ is given by

$$\vec{J}_i(n,:)=\left(G_i\vec{I}_i(n)\right)^t \quad (5)$$

The output symbols are the sum over all encoders:

$$\vec{d}(n,:)=h\left(\bmod\left(\sum_{i=1}^{k}\vec{J}_i(n,:),\ M\right)\right) \quad (6)$$

where the modulo operation is done component wise on the vector. An example is the $L_t=2$, 4-PSK, 8 state, two input bit code from [8]. The trellis is shown in Fig. 3 and the generator matrices are

$$G_1=\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 2 \end{bmatrix},\quad G_2=\begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix} \quad (7)$$

III. SPACE-TIME TCM WITH UNIFORM DISTANCE SPECTRUM

Performance analysis of TCM is greatly simplified when the DS is uniform, because then it is sufficient to consider the DS only from the all-zeros path, rather than from all paths. Linear convolutional codes have a UDS, but unfortunately, bandwidth efficient TCM are not linear [16]. However, numerous researchers have shown that codes with weaker conditions than linearity still have a UDS, as summarized in [17]. To date, no one has shown a method for finding codes where the DS wrt e_P is uniform. This work has found several codes with a UDS wrt e_P, but no general method has been developed. Several of the challenges encountered in this endeavor are noted here.

Geometric uniformity (GU) [18] is a very useful property in single antenna coding theory but it proves less so with multiple antennas. Geometric uniformity is one of the most restrictive types of symmetry that provides a UDS . GU codes were originally defined wrt Euclidean distance, but the concept can be extended to GU wrt e_P or e_H. To aid in discussion we provide several standard definitions.

Definition 1. The Euclidean distance for space-time codes is $\|D_{\alpha\beta}\|_E \triangleq \sqrt{\sum_{n=1}^{N_c}\sum_{l=1}^{L_t}|d_{\alpha\beta}(n,l)|^2}$.

Definition 2. An ***isometry*** u of complex Euclidean N-space $\mathbb{C}^N$ is a transformation $u\colon \mathbb{C}^N \to \mathbb{C}^N$ that preserves Euclidean distances,

$$\|u(x)-u(y)\|_E=\|x-y\|_E,\quad \forall\, x,y\in\mathbb{C}^N \quad (8)$$

where $u(x)$ and $u(y)$ are the images of x and y under the transformation u.

This will be called an isometry wrt Euclidean distance. Similarly, an isometry wrt e_P or e_H is a transformation of a complex space $u\colon \mathbb{C}^{N_c\times L_t}\to\mathbb{C}^{N_c\times L_t}$ that preserves e_P or e_H, respectively. For Euclidean distance all isometries are composed of translations and unitary transformations of the form [18]

$$u_{R,\tau}(x)=Rx+\tau \quad (9)$$

Translations and unitary transformations are also isometries wrt both e_P and e_H [11]. Hence $u_{R,\tau}$ is an isometry wrt both e_P and e_H when $x\in\mathbb{C}^{N_c\times L_t}$, $R\in\mathbb{C}^{N_c\times N_c}$ is unitary, $R^HR=I$ and τ is an arbitrary element of $\mathbb{C}^{N_c\times L_t}$.

Definition 3 [18]. A signal set S is ***geometrically uniform*** when given any two points s and s' in S, there exists an isometry $u_{s,s'}$ that transforms s to $s's'$ while leaving S invariant:

$$\begin{aligned} u_{s,s'}(s)&=s', \\ u_{s,s'}(S)&=S \end{aligned} \quad (10)$$

The signal set is GU wrt the distance measure the isometry preserves; i.e. if the isometry is wrt e_P, the signal set is GU wrt e_P. The signal set could be a constellation or an entire code. Theorem 1 in [18] proves that GU codes have a UDS. It is also stated in [18] that for GU codes with PSK constellations, the τ in (9) must equal zero.

Theorem 1. FFC TCM that are geometrically uniform wrt e_P and use PSK constellations achieve at most one level of diversity.

Proof:

Let D_α be the all-zeros codeword, which has rank one because FFC codes have an all-zeros codeword matrix with all rows identical. Since the constellation is PSK, the isometry must be of the form $u_{R,\tau}(x)=Rx$. By Definition 2, for every codeword D_β there exists a unitary matrix R such that $D_\beta=RD_\alpha$. Thus $D_{\alpha\beta}=D_\alpha-D_\beta=(I-R)D_\alpha$. From linear algebra, the rank of a product of two matrices is less than or equal to the minimum of their ranks. Hence $\operatorname{rank}(D_{\alpha\beta})=e_H(D_{\alpha\beta})\leq 1$, and at most one level of diversity can be achieved. □

The possible ways to overcome this problem are not appealing. There are trellis codes that don't have a transition from state zero back to state zero, and these can have an all-zeros codeword with rank greater than one. However, this type of code is not widely used and is more difficult to work with. Another potential solution is to use non-PSK constellations.

Alternatively, a less restrictive condition than GU must be used to obtain codes with a UDS. Even this is difficult. Exhaustive computer search has shown that for $L_t=3$, $k=1$ and BPSK constellations, no four or eight state FFC TCM codes with a UDS wrt e_P achieve more than one level of diversity. The most likely path of future research is to develop codes that have a DS that is "close enough" to uniform.

The authors in [8] claim to have designed a GU space-time code which achieves full diversity. Their 4 state, 4-PSK code (in Fig. 5) does achieve full diversity and is GU wrt Euclidean distance, but it does not have a UDS wrt e_P, so it can not be GU wrt e_P. The isometry used in [8] is wrt Euclidean distance; the product distance (or determinant criteria) was not addressed in their discussion of geometric uniformity.

The EEPs highlighted in Fig. 5 show that AT&T's code does not have a UDS wrt e_P. For the EEP from the all-zeros path shown in Fig. 5a, $e_P=80$. When a code has a UDS, the DS from every transmitted path is the same. That implies that for two transmitted codewords D_α and $\tilde{D}_\alpha$, every difference codeword produced by a D_β will correspond to a difference codeword produced by a $\tilde{D}_\beta$ with the same e_P. An EEP from a different transmitted path is shown in Fig. 5b. This EEP has $e_P=144$. There are no length 4 EEPs from the all-zeros path with $e_P=144$, so the code is clearly not GU wrt e_P. This EEP is actually better than the EEP from the all-zeros path. However, there are 4 state, 4 PSK codes that are GU wrt Euclidean distance, but $e_{P\min}$ from the all-zeros path is not as bad as the worst e_P from some other transmitted path. For example, the code

$$G_1=\begin{bmatrix}0 & 2\\ 2 & 0\end{bmatrix},\quad G_2=\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix} \tag{11}$$

also has $e_{P\min}=4$, but the worst EEP when the all-zeros codeword is transmitted has $e_P=12$.

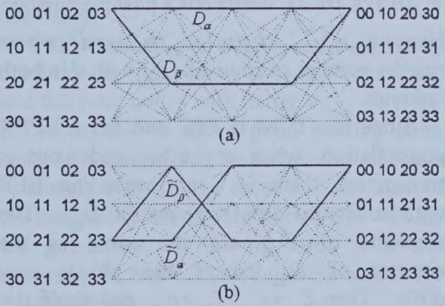


Fig. 5. Two EEPs which prove AT&T's 4 state, 4 PSK code is not GU wrt e_P.

The following theorem justifies using the UDS analysis of the TCM in some cases where the DS is not uniform. It shows that in certain cases, the DS wrt e_P from any codeword is at least as good as the DS wrt e_P from the all-zeros codeword. This suggests that codes without a UDS may exist that are better than the best codes with a UDS.

Theorem 2. If an $L_t=2$, full diversity FFC TCM has a BPSK constellation, every EEP from the all zeros path has a corresponding EEP from some other transmitted path with the same or better e_P.

Proof:

For L_t=2 the codeword difference matrix of an EEP may be written $D_{\alpha\beta} = [\vec{d}_{\alpha\beta}(:,1)\ \vec{d}_{\alpha\beta}(:,2)]$ where $\vec{d}_{\alpha\beta}(:,l)$, $l=1{:}L_t$ is an $N_c \times 1$ vector with elements $d_{\alpha\beta}(n,l)$, $n=1{:}N_c$. Since the constellation is BPSK, the effective product distance depends on

$$C_2 = D^t_{\alpha\beta} D_{\alpha\beta} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \qquad \textbf{(12)}$$

where $c_{l_1 l_2} = \vec{d}^{\,t}_{\alpha\beta}(:,l_1)\vec{d}_{\alpha\beta}(:,l_2)$ is the correlation of the column vectors. It is clear that the autocorrelations c_{11} and c_{22} do not depend on the signs of the elements of $D_{\alpha\beta}$ and that $c_{12}=c_{21}$. Since by hypothesis the TCM is full rank, $e_P(D_{\alpha\beta}) = \det(C_2) = c_{11}c_{22} - c_{12}^2$, and the sign of c_{12} is not important. Without loss of generality assume that the symbols of the all-zeros codeword are positive. Then when the all-zeros codeword is transmitted the signs of the elements of $D_{\alpha\beta}$ are all the same, so $d_{\alpha\beta}(n,l) \in \{0,d\}$, where d is a positive number. Hence $c_{12} = \sum_{n=1}^{N_c} d_{\alpha\beta}(n,1)d_{\alpha\beta}(n,2) = rd^2$, where r is the number of rows of $D_{\alpha\beta}$ with d in both columns.

Since the TCM is FFC and has a BPSK constellation, when any other codeword is transmitted there is a corresponding EEP with codeword difference matrix $\tilde{D}_{\alpha\beta}$. The elements of $\tilde{D}_{\alpha\beta}$ are the same as the elements of $D_{\alpha\beta}$ except the phases may differ. Then $\tilde{c}_{11} = c_{11}$, $\tilde{c}_{22} = c_{22}$, and $\mathrm{abs}(\tilde{c}_{12}) = \mathrm{abs}(\sum_{n=1}^{N_c} (\pm d_{\alpha\beta}(n,1))(\pm d_{\alpha\beta}(n,2))) \le c_{12}$. Since $e_P(\tilde{D}_{\alpha\beta}) = \det(\tilde{C}_2) = \tilde{c}_{11}\tilde{c}_{22} - \tilde{c}_{12}^2$, the DS from the other codeword is at least as good as the DS from the all-zeros codeword. □

Theorems 1 and 2 apply not only to FFC TCM, but any TCM that has a transition from state zero back to state zero. For the more general TCM, theorem 2 only holds if the TCM is GU wrt Euclidean distance. Theorem 2 does not always hold for larger constellations or higher levels of diversity.

An example non-UDS wrt e_P code with a better DS than the best possible UDS wrt e_P code is shown in Fig. 6. Code A is the best possible L_t=2, k=v=m=1 FFC TCM with a UDS wrt e_P, but its performance is slightly worse than code B, which does not have a UDS wrt e_P. Both codes are GU wrt Euclidean distance. Both codes have $e_{H\min} = 2$ and $e_{P\min} = 4$, but some of the longer EEPs of code B have better e_P than code A when codewords other than the all-zeros codeword are transmitted. Furthermore, code A is not GU by Definition 2 because no isometries of the form in (9) preserve the code. These claims are proven in [11].

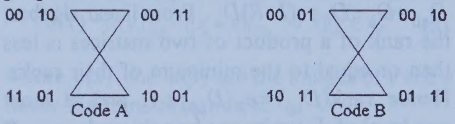


Fig. 6. Two TCM codes. Code A has a UDS wrt e_P and code B does not, but the performance of code B is slightly better.

In summary, it is not as easy to design codes with a UDS wrt e_P as it is for Euclidean distance. Full diversity PSK GU wrt e_P codes do not exist. Also, Theorem 2 shows that codes with a UDS may not be optimal. Without a UDS, performance analysis is extremely computationally intensive, so finding some notion of linearity that applies to TCM for transmitter diversity is an important topic of future research. The most hopeful fact is that Theorem 2 implies the existence of codes where the DS from any codeword is at least as good as the DS from the all-zeros codeword. For these codes, an upper bound of the BEP can still be made using only the DS from the all-zeros codeword. Many codes exist where $e_{P\min}$ is the same for every transmitted path.

IV. TRELLIS CODE DESIGN FOR TRANSMITTER DIVERSITY

This section discusses and generalizes the design rules proposed in [8]. Also a repetition coded delay diversity scheme is proposed which uses the fewest states possible to achieve full diversity. These results serve as a lower bound on $e_{P\min}$ for a given throughput and diversity level. Finally, a search procedure is used to find the best codes.

A. AT&T Design Rules (L_t=2)

In [8] the researchers at AT&T proposed the following two design rules to guarantee trellis codes with two levels of diversity.

Design Rule 1: Transitions departing from the same state differ in the first[1] symbol.

Design Rule 2: Transitions arriving at the same state differ in the second[1] symbol.

As stated, these rules are not sufficient to guarantee full diversity. The code in Fig. 7 meets these design rules but the highlighted EEP is not full rank.

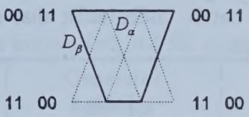


Fig. 7. A two state BPSK TCM which meets AT&T's design rules but does not achieve full diversity.

The codeword difference matrix of the highlighted EEP is clearly not full rank. However, this is a minor technicality. It is implicit in these rules that transitions departing from the same state are the same in the second symbol, and transitions arriving at the same state are the same in the first symbol. When this is followed, the codeword difference matrices are always full rank and have the form

$$D_{\alpha\beta} = D_\alpha - D_\beta = \begin{bmatrix} y(1) & 0 \\ \vdots & \vdots \\ 0 & y(2) \end{bmatrix} \tag{13}$$

where $y(1)$, $y(2) \in \mathbb{C}$ are nonzero values and the unspecified part of the matrix is arbitrary.

These rules are not necessary to achieve full diversity, and in some cases they rule out optimal codes. Still, they are a good starting point because of their simplicity, and because they make it possible to hand design many codes which turn out to be optimal.

B. Zeros Symmetry ($L_t \geq 2$)

The AT&T rules can be extended to any level of diversity, L_t, by constraining the codeword difference matrices to be both upper- and lower-triangular, as illustrated in (14) for L_t=3. This will be referred to as zeros symmetry (ZS).

$$D_{\alpha\beta} = \begin{bmatrix} y(1) & 0 & 0 \\ x(2) & y(2) & 0 \\ \vdots & x(3) & x(1) \\ x(5) & \vdots & x(4) \\ x(8) & x(6) & \vdots \\ 0 & y(3) & x(7) \\ 0 & 0 & y(4) \end{bmatrix} \tag{14}$$

The $y(n) \in \mathbb{C}$ can be any nonzero value and the $x(n) \in \mathbb{C}$ can be any value. This restriction is sufficient for full rank, but clearly not necessary.

Unfortunately, since ZS is not a necessary condition for full rank it is overly restrictive, and full rank codes with better $e_{P\min}$ do exist. However, this is still a good starting point because it greatly reduces the number of codes to search, and in many cases it does result in the best code. In addition, since TCM with ZS are guaranteed to be full rank, it is not necessary to test each code in the search for full diversity.

Delay diversity [12] is a special case of zeros symmetry. Delayed copies of the transmitted symbol are sent consecutively across the L_t antennas, instead of all in the same symbol interval. Trellis coded delay diversity was demonstrated in [8] for L_t=2.

Repetition coded delay diversity TCM may be extended to any level of diversity, simply by using $G_i = 2^{i-1} I$, $i = 1{:}k$, where I is an $L_t \times L_t$ identity matrix. This means that the number of memory elements per information bit is $v_i = L_t - 1$ and the total number of memory elements is $v = k(L_t - 1)$.

Theorem 3. This delay diversity scheme uses the fewest states possible to achieve full diversity.

---

[1] In the AT&T paper, the words "first" and "second" in rules 1 & 2 are reversed, but both ways are equivalent. Here the rules are stated so the delay diversity codes are on the forward diagonal of the codeword matrix instead of the backward diagonal.

Proof:

The standard Hamming distance of any trellis code is bounded by $d_H \le \lfloor v/k \rfloor + 1$. To achieve full diversity, it is required that $L_t \le d_H$, thus

$$\begin{gathered} L_t \le \lfloor v/k \rfloor + 1 \\ kL_t \le k\lfloor v/k \rfloor + k \le k(v/k) + k \\ v \ge k(L_t - 1) \end{gathered} \tag{15}$$

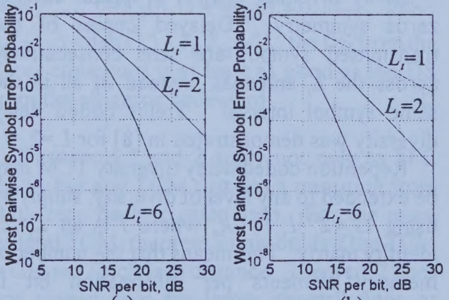
Since delay diversity TCM has $v = k(L_t - 1)$, it uses the fewest states possible. □

The worst case product distance for repetition coded delay diversity occurs on the shortest error event path from the all-zeros path, where the code symbols of the two paths are nearest neighbors in the constellation. Thus

$$e_{P\min} = \Delta^{2L_t} \tag{16}$$

where Δ is the minimum Euclidean distance between constellation points. Hence IQ TCM has the same error performance of PAM, but the throughput is doubled. The worst case exact pairwise SEP is displayed in Fig. 8. It may be seen that 16-IQ TCM is about 4 dB better than 16-PSK for this coding scheme.

For some of these codes, the worst case e_P could be improved by using a better code than a repetition code, as was done in [8] for L_t=2. This tends to decrease e_P of the longer error event paths, so the average BEP is not improved by as much as might at first be expected.



(a) (b)

Fig. 8. Worst case SEP for delay diversity, L_t=1:6. (a) *k*=4 (16-PSK) and (b) *k*=2 (4-PAM) or *k*=4 (16-IQ).

C. Search for More Powerful Codes

A code search based on the principles discussed so far was performed; the details of the search algorithm are in [11]. Generator matrices for some of the best known codes are presented in Tables I and II, and more codes are presented in [11]. The worst case SEPs for the *k*=1 codes are shown in Fig. 9. Note that for *k*=1 no minimal complexity codes were found with better performance than the repetition coded delay diversity TCM. This is not true for *k*>1, as seen in Tables II and III. Also, the codes in Tables II and III show that AT&T design rules are not optimal for *k*>2, because better codes exist that don't satisfy their rules.

Table I

Best known *k*=1, *m*=1 codes

L_t	v	Generator Matrix	$e_{P\min}$
2	1	$G_1 = \begin{bmatrix} 10 \\ 01 \end{bmatrix}$	4.000
2	6	$G_1 = \begin{bmatrix} 1101110 \\ 0111101 \end{bmatrix}$	56.000
3	2	$G_1 = \begin{bmatrix} 100 \\ 010 \\ 001 \end{bmatrix}$	2.370
3	6	$G_1 = \begin{bmatrix} 1101100 \\ 0111010 \\ 0011111 \end{bmatrix}$	56.889
5	4	$G_1 = \begin{bmatrix} 10000 \\ 01000 \\ 00100 \\ 00010 \\ 00001 \end{bmatrix}$	0.328
5	6	$G_1 = \begin{bmatrix} 1100000 \\ 0101000 \\ 0010100 \\ 0001110 \\ 0000011 \end{bmatrix}$	5.243

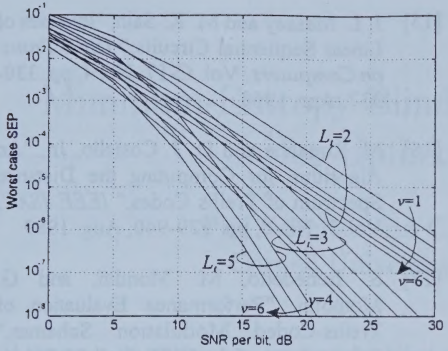


Fig. 9. Worst case SEP for the L_t=2, 3, and 5 codes from Table I.

Table II
Best known k=2, 4-PSK codes

L_t	v	ZS	Generator Matrices	$e_{P\min}$
2	2	yes	$G_1=\begin{bmatrix}10\\01\end{bmatrix}, G_2=\begin{bmatrix}20\\02\end{bmatrix}$	4.000
2	2	no	$G_1=\begin{bmatrix}10\\02\end{bmatrix}, G_2=\begin{bmatrix}02\\12\end{bmatrix}$	8.000
2	3	yes	$G_1=\begin{bmatrix}130\\011\end{bmatrix}, G_2=\begin{bmatrix}20\\02\end{bmatrix}$	12.000
2	3	no	$G_1=\begin{bmatrix}102\\012\end{bmatrix}, G_2=\begin{bmatrix}12\\20\end{bmatrix}$	16.000

Table III
Best known k=2, 4-PAM or 16-IQ codes

L_t	v	ZS	Generator Matrices	$e_{P\min}$
2	2	yes	$G_1=\begin{bmatrix}10\\01\end{bmatrix}, G_2=\begin{bmatrix}20\\02\end{bmatrix}$	0.640
2	2	no	$G_1=\begin{bmatrix}10\\02\end{bmatrix}, G_2=\begin{bmatrix}02\\12\end{bmatrix}$	2.560
2	3	yes	$G_1=\begin{bmatrix}210\\001\end{bmatrix}, G_2=\begin{bmatrix}10\\02\end{bmatrix}$	2.560
2	3	no	$G_1=\begin{bmatrix}221\\132\end{bmatrix}, G_2=\begin{bmatrix}02\\20\end{bmatrix}$	7.040

V. CONCLUSION

We proved that GU wrt e_P FFC TCM with PSK constellations cannot achieve more than one diversity level. Additionally it was shown that codes without a UDS wrt e_P can outperform the best codes with a UDS wrt e_P. An important topic of future research is to find some notion of linearity that applies to space-time TCM, since this would greatly reduce the complexity of performance analysis.

Design rules to guarantee full diversity for any desired diversity level were proposed. It was shown that delay diversity satisfies these rules, and a delay diversity scheme that can achieve any desired throughput and diversity level with the minimal number of encoder states was presented. A search routine was used to find codes with more memory that have better error performance.

REFERENCES

[1] G. Ungerboeck, "Channel Coding with Multilevel/Phase Signals", *IEEE Trans. on Info. Theory*, Vol. IT-28, No. 1, pp. 55-67, Jan. 1982.

[2] D. Divsalar and M. K. Simon, "The Design of Trellis Coded MPSK for Fading Channels: Performance Criteria," *IEEE Trans. Com.*, Vol. 36, No. 9, pp. 1004-1012, Sep. 1988.

[3] C.-E. W. Sundberg, N. Seshadri, "Coded Modulations for Fading Channels: An Overview," *European Trans. Telecom.*, Vol. 4, No. 3, pp. 309-323, May-June 1993.

[4] A. J. Viterbi, J. K. Wolf, E. Zehavi, and R. Padovani, "A Pragmatic Approach to Trellis-Coded Modulation," *IEEE Com. Mag.*, pp. 11-19, July 1989.

[5] P. K. Ho, J. K. Cavers and J. L. Varaldi, "The effects of Constellation Density on Trellis Coded Modulation in Fading Channels," *IEEE Trans. on Vehic. Tech.*, Vol. 42, No. 3, pp. 318-325, Aug. 1993.

[6] S. A. Al-Semari and T. E. Fuja, "I-Q TCM: Reliable Communication Over the

Rayleigh Fading Channel Close to the Cutoff Rate," *IEEE Trans. on Info. Theory*, Vol. 43, No. 1, pp. 250-262, Jan. 1997.

[7] W. C. Jakes, "A Comparison of Specific Space Diversity Techniques for Reduction of Fast Fading in UHF Mobile Radio Systems," *IEEE Trans. on Vehic. Tech.*, Vol. VT-20, No. 4, pp. 81-92, Nov. 1971.

[8] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction," *IEEE Trans. on Info. Theory*, Vol. 44, No. 2, pp. 744-765, March 1998.

[9] J.-C. Guey, M. P. Fitz, M. R. Bell and W.-Y. Kuo, "Signal Design for Transmitter Diversity Wireless Communication Systems over Rayleigh Fading Channels," *IEEE VTC '96*, pp. 136-140.

[10] J.-C. Guey, M. P. Fitz, M. R. Bell and W.-Y. Kuo, "Signal Design for Transmitter Diversity Wireless Communication Systems over Rayleigh Fading Channels," accepted by *IEEE Trans. Com.*

[11] J. Grimm, "Transmitter Diversity Code Design for Achieving Full Diversity on Rayleigh Fading Channels," Ph.D. Dissertation, Purdue University, Dec. 1998.

[12] N. Seshadri and J. H. Winters, "Two Signaling Schemes for Improving the Error Performance of Frequency Division Duplex (FDD) Transmission Systems Using Transmitter Antenna Diversity", *Int'l J. of Wireless Info. Networks*, Vol. 1, No. 1, pp. 49-60, 1994.

[13] D. Divsalar and M. K. Simon, "Multiple Trellis Coded Modulation (MTCM)," *IEEE Trans. on Comm.*, Vol. COM-36, pp. 410-419, Apr. 1988.

[14] R. van Nobelen and D. P. Taylor, "Geometrically Uniform Trellis Codes for the Rayleigh Fading Channel," *GLOBECOM*, pp. 1820-1824, 1995.

[15] J. L. Massey and M. K. Sain, "Inverses of Linear Sequential Circuits," *IEEE Trans. on Computers*, Vol. C-17, No. 4, pp. 330-337, Apr. 1968.

[16] M. Rouanne and D. J. Costello, Jr., "An Algorithm for Computing the Distance Spectrum of Trellis Codes," *IEEE JSAC*, Vol. 7, No. 6, pp. 929-940, Aug. 1989.

[17] S. Benedetto, M. Mondin, and G. Montorsi, "Performance Evaluation of Trellis-Coded Modulation Schemes," *Proceedings of the IEEE*, Vol. 82, No. 6, pp. 832-855, June 1994.

[18] G. D. Forney, Jr., "Geometrically Uniform Codes," *IEEE Trans. on Info. Theory*, Vol. 37, No. 5, pp. 1241-1260, Sep. 1991.

Minimum-Entropy Blind Acquisition/Equalization for Uplink DS-CDMA

Phil Schniter and C. Richard Johnson, Jr.*
Cornell University, Ithaca, NY 14853.

Abstract: In this paper, we consider blind estimation of linear chip-spaced receivers for the demodulation of a particular short-code DS-CDMA mobile user under multipath propagation and in the absence of timing information. We propose a family of schemes for blind acquisition and equalization based on Donoho's Minimum Entropy principle and propose a specific algorithm that uses the second- and fourth-order moments of a pre-whitened chip-rate received signal. The proposed algorithm can be considered a near-far resistant initialization procedure for, and application of, the Constant Modulus Algorithm (CMA) to DS-CDMA.

1 Introduction

Direct sequence code division multiple access (DS-CDMA) systems have received considerable attention as a flexible means of communication between multiple mobile users and centralized base stations. Since CDMA users share the same time and frequency resources, demodulation of a particular user is principally concerned with suppression of interference from other users. The link from base station to mobiles is referred to as the "downlink" and is typically characterized by synchronous equal-power data transmission to all users. More challenging for demodulation is the "uplink," from mobiles to base station, where different user transmissions are typically asynchronous and of widely disparate power levels. Reliable demodulation might also require mitigation of multipath interference, especially in recently proposed "wideband" CDMA schemes (e.g., IMT-2000) where multipath effects can be significant.

In this paper, we consider blind estimation of linear receivers for demodulation of a (possibly weak) asynchronous CDMA user under multipath propagation. By "blind," we mean that all interfering code vectors, all user timings, and all multipath channels are unknown, and that no training information is available. Thus, obtaining reliable symbol estimates involves both blind equalization *and* blind timing acquisition of the desired user. Previously proposed schemes rely solely on the second-order statistics (SOS) of the received data and are based on either subspace-constrained optimization [1, 2] or subspace decomposition [3]. The use of constraints implies that the minimizers of such criteria are *not*, in general, global MSE minimizers. Proponents claim that "this price is paid in exchange for the advantage of not having a training sequence" [2, p. 107], but must we really pay this price? Similarly, subspace techniques assume that the signal and noise subspaces can be accurately estimated, but is this usually possible?

---

*Supported in part by NSF grant MIP-9509011, a Schlumberger Foundation Fellowship, and Applied Signal Technology. Direct all correspondence to Phil Schniter: `schniter@ee.cornell.edu`.

In this paper, we propose a family of blind acquisition/equalization schemes that make inherent use of the higher-order statistics (HOS) of the received signal. The central concept is rooted in Donoho's Minimum Entropy (ME) principle [4], a fundamentally different approach to blind acquisition/equalization than [1, 2, 3]. A major advantage of the minimum entropy techniques is that they are capable of achieving *global* MMSE solutions under ideal conditions (to be discussed) while exhibiting robust performance in realistic (non-ideal) conditions. We present a specific ME algorithm that uses the second- and fourth-order moments of a whitened chip-rate version of the received signal. It can be considered a near-far resistant application of CMA to uplink DS-CDMA in the absence of timing information.

Notation: We use bold lowercase for vectors, bold uppercase for matrices, $(\cdot)^t$ for transposition, $(\cdot)^*$ for conjugation, and $(\cdot)^H$ for conjugate transposition. All vector/matrix indices start with zero (i.e., we denote the first element of $\mathbf{x}$ by x_0 rather than x_1), and n used as a symbol-rate index, i as a chip-rate index, k as a user index, and ν as a hypothesis index. Finally, $\lceil\cdot\rceil$ denotes ceiling, $\langle\cdot\rangle_N$ modulo-N, $\mathrm{E}\{\cdot\}$ expectation, $*$ convolution, $\|\mathbf{x}\|_p := \sqrt[p]{\sum_m |x_m|^p}$ the ℓ_p norm of $\mathbf{x}$, and $\mathbf{e}_m$ a vector with a one in the m^{th} position and zeros elsewhere.

2 DS-CDMA System Model

We formulate the uplink DS-CDMA model as follows. K mobile users transmit simultaneously to a base station, where the k^{th} user transmits a symbol (i.e., "bit") sequence $\{s_n^{(k)}\}$ indexed by n and with symbol spacing of T seconds. The users are each assigned a "short" spreading code composed of N chips with chip duration $T_c = T/N$. We shall represent the k^{th} user's code by the vector $\mathbf{c}^{(k)} := (c_0^{(k)}, \ldots, c_{N-1}^{(k)})^t$. A given user's short code is multiplied by each of his symbols prior to transmission, so that the transmitted data sequence is spaced at the chip rate.

Now consider the baseband model illustrated in Figure 1. Each mobile transmits his chip-rate sequence through a pulse shaping filter with impulse response $p(t)$. In typical applications, $p(t)$ is bandlimited to approximately $1/T_c$ using, e.g., a square-root raised-cosine filter. Before reaching the receiver, each transmitted waveform passes through a user-specific atmospheric propagation channel which includes, e.g., multipath and path-loss effects. We restrict our focus to a linear time-invariant approximation to the propagation channel and denote the resulting impulse response by $g^{(k)}(t)$. In addition to linear distortion, the propagation channel contributes additive noise $n(t)$.

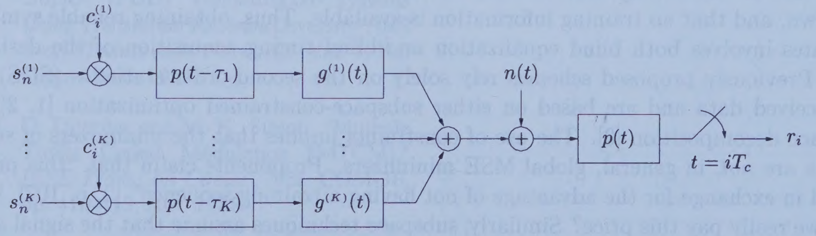


Figure 1: DS-CDMA system model with chip-sampling receiver.

2.1 Discrete Model

The receiver sees the sum of all transmitted signals plus noise and samples the output of a single "chip-matched" filter at the chip rate $1/T_c$. As is typical, we assume that the chip-matched filter is identical to the transmit filter $p(t)$. User-specific time delays $\{\tau_k\}$ are used to capture the lack of timing coordination, or asynchronicity, among users. Without loss of generality[1], time delay is assumed in the range $0 \le \tau_k < T$.

The model of Figure 1 admits the discrete equivalent of Figure 2, where $\uparrow N$ denotes upsampling and $\overleftarrow{\mathbf{c}}^{(k)} := (c_{N-1}^{(k)}, \dots, c_0^{(k)})^t$ the reverse-ordered code vector, and where $\mathbf{h}^{(k)} := (h_0^{(k)}, \dots, h_{L_h-1}^{(k)})^t$ contains the T_c-spaced impulse response samples of the convolved transmit filter, propagation channel, and receiver filter:

$$h_i^{(k)} \;=\; p(t-\tau_k) * g^{(k)}(t) * p(t)\big|_{t=iT_c}. \tag{1}$$

In (1) we assumed that $p(t)$ and $g^{(k)}(t)$ are causal and time-limited so that, for all k, $p(t-\tau_k) * g^{(k)}(t) * p(t)|_{t=iT_c} = 0$ for $i \ge L_h$ and $i < 0$. Note that using one number, L_h, to characterize the length of every user's discrete channel implies that some of the $\mathbf{h}^{(k)}$ may be zero-padded. Finally, note that the sampled noise process $\{w_i\}$ in Figure 2 is given by $w_i = n(t) * p(t)|_{t=iT_c}$.

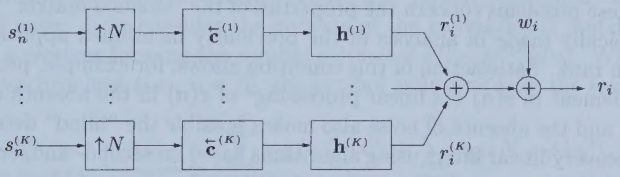


Figure 2: Discrete-time equivalent DS-CDMA system model with chip-sampling receiver.

From now on we restrict our attention to finite-length linear receivers. Collecting, say, the L_r most recent samples at time n in the vector $\mathbf{r}(n)$, it is possible to express $\mathbf{r}(n)$ as a linear combination of transmitted symbols. To see this, we first write $\mathbf{r}(n)$ in terms of specific user contributions plus noise:

$$\mathbf{r}(n) \;=\; \sum_{k=1}^{K} \mathbf{r}^{(k)}(n) + \mathbf{w}(n), \tag{2}$$

where $\mathbf{r}^{(k)}(n)$ is a vector composed of the L_r most recent noiseless chip-rate k^{th} user contributions $r_i^{(k)}$, and $\mathbf{w}(n)$ is composed of the L_r most recent chip-rate noise samples w_i (see Figure 2). The k^{th} user contribution can be written as a matrix-vector product:

$$\mathbf{r}^{(k)}(n) \;=\; \mathbf{H}^{(k)} \mathbf{C}^{(k)} \mathbf{s}^{(k)}(n), \tag{3}$$

where $\mathbf{H}^{(k)}$ is an $L_r \times L_r + L_h - 1$ Toeplitz matrix and $\mathbf{C}^{(k)}$ is a $L_r + L_h - 1 \times \lceil \frac{L_r + L_h - 1}{N} \rceil$ matrix structured as below,

$$\mathbf{H}^{(k)} := \begin{pmatrix} h_0^{(k)} & h_1^{(k)} & \dots & h_{L_h-1}^{(k)} & & & \\ & h_0^{(k)} & h_1^{(k)} & \dots & h_{L_h-1}^{(k)} & & \\ & & \ddots & & & \ddots & \\ & & & h_0^{(k)} & h_1^{(k)} & \dots & h_{L_h-1}^{(k)} \end{pmatrix}, \mathbf{C}^{(k)} := \begin{pmatrix} \mathbf{c}^{(k)} & & & \\ & \mathbf{c}^{(k)} & & \\ & & \ddots & \\ & & & \mathbf{c}^{(k)} \end{pmatrix},$$

---

[1] Assuming stationary source processes, the received signal will be cyclo-stationary with period T. Thus, the statistics of the received sequence will be insensitive to shifts of nT in user delay.

but with the rightmost code vector possibly truncated to length $\langle L_r+L_h-2\rangle_N+1$. The vector of the k^{th} user's source symbols influencing $\mathbf{r}(n)$, namely $\mathbf{s}^{(k)}(n) := (s_n^{(k)}, \ldots, s_{n-L_s+1}^{(k)})^t$, is of length $L_s := \lceil \frac{L_r+L_h-1}{N} \rceil$. Equation (2) can then be written

$$\mathbf{r}(n) \;=\; \sum_{k=1}^{K} \mathbf{H}^{(k)}\mathbf{C}^{(k)}\mathbf{s}^{(k)}(n) + \mathbf{w}(n) \;=\; \mathcal{H}\,\mathbf{s}(n) + \mathbf{w}(n), \tag{4}$$

where the $L_r \times KL_s$ multiuser code-channel matrix $\mathcal{H}$ and the $KL_s \times 1$ multiuser source vector $\mathbf{s}(n)$ are defined by

$$\mathcal{H} \;:=\; \left(\mathbf{H}^{(1)}\mathbf{C}^{(1)}, \mathbf{H}^{(2)}\mathbf{C}^{(2)}, \ldots \mathbf{H}^{(K)}\mathbf{C}^{(K)}\right), \tag{5}$$

$$\mathbf{s}(n) \;:=\; \left(\mathbf{s}^{(1)^t}(n), \mathbf{s}^{(2)^t}(n), \ldots, \mathbf{s}^{(K)^t}(n)\right)^t. \tag{6}$$

2.2 Rank of Code-Channel Matrix $\mathcal{H}$

Expression (4) for the received DS-CDMA signal has the same basic structure seen in many multichannel deconvolution problems (e.g., wideband beamforming, wideband source separation, crosstalk suppression, and cross-polarization). The principle differences among these problems concern the properties of the "channel matrix" $\mathcal{H}$. A critical assumption typically made in analysis of the previously mentioned applications is that $\mathcal{H}$ is full column rank. Satisfaction of this condition allows, for example, perfect recovery of any source element in $\mathbf{s}(n)$ via linear processing[2] of $\mathbf{r}(n)$ in the absence of noise. Full column rank $\mathcal{H}$ and the absence of noise also makes possible the "blind" determination of these perfect-recovery linear filters using algorithms based on second- and/or higher-order statistics. Non full-column rank $\mathcal{H}$ can degrade the performance of blind algorithms, in some cases catastrophically. These reasons motivate a closer look at the rank of $\mathcal{H}$ for the specific case of asynchronous DS-CDMA under multipath propagation.

Let us consider the dimensional requirements necessary for full column rank $\mathcal{H}$. To have at least as many rows as columns, one requires $L_r \geq K\lceil \frac{L_r+L_h-1}{N} \rceil$. In the case that $K = N$, the requirement becomes $L_r \geq L_r+L_h-1+C$, where C is an integer in the range $0 \leq C < N$. It follows directly that a necessary condition for full column rank $\mathcal{H}$ with $K = N$ users is $L_h = 1$, or, in other words, the absence of asynchronism or multipath! In the case $K < N$, it can be shown that choosing $L_r > \lceil \frac{K(N+L_h-1)-N}{N-K} \rceil$ guarantees that the dimensional requirement is satisfied. To summarize, for a single-sensor DS-CDMA system with $\geq N$ users, asynchronism or multipath will ensure that $\mathcal{H}$ is *not* full column rank for any finite observation length L_r.

Practical cellular communication systems are based on the use of multiple cells, each cell composed of one base station and $\leq N$ mobiles. One of the principle advantages of CDMA in the cellular context is its potential for universal frequency reuse. This implies, however, that other-cell signals will interfere with in-cell communication. In fact, it has been suggested that the typical out-of-cell interference level is roughly half of the in-cell interference level [5]. Thus, a multi-cell model should incorporate $K \gg N$ significant users and, by the arguments above, will result in non full-column-rank $\mathcal{H}$. We note that the rank of $\mathcal{H}$ may be increased through the use of a receiver with multiple sensors [3], but the number required to achieve full rank may be prohibitively large.

---

[2] Perfect linear recovery follows directly from the existence of a left inverse for full column rank $\mathcal{H}$.

3 Blind Acquisition/Equalization

With knowledge of only the ℓ^{th} user's code $\mathbf{c}^{(\ell)}$ and the received sequence $\{r_i\}$, we consider the problem of linear estimation of the ℓ^{th} user's symbol sequence at the base station (up to arbitrary phase and symbol delay). Specifically, we generate linear symbol estimates $\hat{s}^{(\ell)}_n = \mathbf{f}^H\mathbf{r}(n)$ using receiver $\mathbf{f}$ and desire that $\hat{s}^{(\ell)}_n = \alpha_\ell s^{(\ell)}_{n-\delta_\ell}$ $\forall n$, where $|\alpha_\ell| = 1$ and $\delta_\ell \in \mathbb{Z}^+$. This is often referred to as the "blind acquisition and demodulation" problem [1] because the channels $\{\mathbf{h}^{(k)}\}$, user delays $\{\tau_k\}$, and interfering users' codes $\{\mathbf{c}^{(k)} : k \neq \ell\}$ are unknown[3] and no training sequences are transmitted.

3.1 Minimum-Entropy Acquisition

The proposed scheme can be motivated, at least on a conceptual level, by (i) the non-Gaussian nature of typical DS-CDMA sources (e.g., BPSK), and (ii) the approximate orthogonality properties of good codes, e.g. Gold codes. By "good" codes we mean those whose cross-correlation coefficients $r_{\ell,k}(i) := \sum_m c^{(\ell)}_m c^{(k)*}_{m-i}$ corresponding to user/timing mismatches are significantly smaller than those corresponding to the correct user/timing combination, i.e., $r_{\ell,\ell}(0) \gg |r_{\ell,k}(i)|$ $\forall i, \ell, k \neq \ell$. For simplicity, let us first consider a situation with chip-level synchronization, equal user powers, zero inter-chip interference (ICI), and no noise. Now consider the outputs of the conventional matched-filter (MF) receiver, i.e., a receiver forming the symbol estimates $\hat{s}^{(\ell)}_n = \mathbf{c}^{(\ell)H}\mathbf{r}(n)$. When the desired user is time synchronized (i.e., $\tau_\ell = 0$), the symbol estimates take the form

$$\hat{s}^{(\ell)}_n = \underbrace{r_{\ell,\ell}(0)s^{(\ell)}_n}_{\text{non-Gaussian}} + \underbrace{\sum_{k\neq\ell}^{K} r_{\ell,k}(-\tfrac{\tau_k}{T_c})s^{(k)}_n + \sum_{k\neq\ell}^{K} r_{\ell,k}(N-\tfrac{\tau_k}{T_c})s^{(k)}_{n-1}}_{\text{approximately Gaussian}}. \tag{7}$$

Assuming non-Gaussian source processes and good codes, the synchronized symbol estimates are a sum of two processes: (i) a non-Gaussian desired process, and (ii) an interference process, which tends towards Gaussianity as K becomes large. When, instead, the MF receiver is *not* synchronized with the desired user, the unsynchronized symbol estimates will take the form of the latter terms in (7), i.e., approximately Gaussian. As described in Donoho's seminal paper [4], the principle of "maximizing distance from Gaussianity" is a fundamental concept that lies at the base of many blind estimation techniques that make use of higher-order statistics. Henceforth we refer to such techniques as "minimum entropy" (ME) methods in the spirit of Donoho, who demonstrated formal connections to Shannon Entropy [4].

The observations above suggest a blind timing acquisition scheme whereby the symbol-rate output streams of delay-hypothesized matched filters are examined for their "degree of Gaussianity." The delay hypothesis leading to the "least Gaussian" stream could be chosen as an estimate of the desired user's delay. The measure of Gaussianity could be constructed in various ways, e.g., using sample-average based estimates of kurtosis.

3.2 Minimum-Entropy Equalization

The previous section focused on the estimation of the desired user's chip delay, i.e., the blind *acquisition* problem. Our goal, however, is really the estimation of desired user symbols in the presence of unknown multipath and interference. Thus, the problem of interest bears close resemblance to a blind *equalization* problem.

---

[3] Though the base station typically demodulates *all* in-cell users, and thus knows more than one code, the assumption of unknown interfering user codes supports a model including out-of-cell users.

Connections to minimum entropy have been established for several popular blind equalization algorithms. For example, the Shalvi-Weinstein (SW) criterion [6] minimizes the kurtosis of the estimates $\hat{s}_n$, defined $K(\hat{s}_n) := \mathrm{E}\{|\hat{s}_n|^4\} - 2\,\mathrm{E}^2\{|\hat{s}_n|^2\} - |\,\mathrm{E}\{\hat{s}_n^2\}|^2$, subject to the power constraint $\mathrm{E}\{|\hat{s}_n|^2\} = \mathrm{E}\{|s_n|^2\}$. Kurtosis is, in fact, a direct measure of "distance from Gaussianity," and it may not be surprising that the SW criterion is nearly identical to Wiggin's ME method discussed in [4]. Furthermore, it has been shown that (in the absence of noise) the minima of the constant modulus (CM) criterion [7] correspond to SW minima, regardless of the rank of $\mathcal{H}$ [8]. Thus, the CM algorithm (CMA) [9] can also be considered a member of the ME family.

A well-known difficulty in applying ME-based algorithms (such as CMA) to multiuser demodulation is that the typical "cost function does not distinguish between desired and interfering symbols, which leads to a number of local minima" [1]. As a partial solution, multiuser versions of CMA have been proposed that consist of parallel single-user demodulators whose outputs are constrained to be mutually uncorrelated (e.g., [10]). Such techniques, however, do not address the problem of locking onto a *specific* user, which is the goal of this paper.

When using an iterative ME equalization technique, the problem of demodulating a specific user reduces to the problem of *properly initializing the blind algorithm.* When the timing of the desired DS-CDMA user is known, the initialization of CMA, for example, appears rather straightforward (see, e.g., [11]). When the desired user timing is unknown, we propose initialization based on blind ME acquisition, as described in Section 3.1. In fact, we consider the goal of blind acquisition not to be the estimation of the desired user's time delay (or group delay in the presence of multipath), but rather the achievement of a good initialization for subsequent blind equalization.

In a power disparate, or "near-far," scenario, the convergence of ME equalization algorithms becomes highly sensitive to choice of initialization, making acquisition of a specific user potentially difficult. For example, the sizes of CMA's regions of convergence (in equalizer parameter space) are proportional to the eigenvalues of the autocorrelation matrix $\mathbf{R}_{rr} := \mathrm{E}\{\mathbf{r}(n)\mathbf{r}^H(n)\}$ [16]. Since weak user transmissions are characterized by signal spaces with small eigenvalues, the size of weak users' convergence regions will also be small and determining a CMA initialization within of one of these small regions may be practically impossible. This phenomenon is demonstrated in Figure 3, where not even initialization at the (global) MMSE solution was adequate for successful convergence of the two weakest users.

In an attempt to mitigate these near-far effects, we propose a whitening of the received signal prior to ME processing. Such "orthogonalization" is commonly used in many blind source separation algorithms [13] as well as some blind equalization algorithms (e.g., [6, 14]). Pre-whitening has the effect of nearly equating each user's region-of-convergence volume, significantly reducing initialization sensitivity.

4 A CMA-based Acquisition/Equalization Algorithm

A CMA-based acquisition/equalization algorithm is presented below:

1) Obtain estimate $\hat{\mathbf{T}}$ of the $L_r \times L_r$ whitening matrix $\mathbf{T} = \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{V}^H$, where $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$ is the eigendecomposition of the received autocorrelation matrix $\mathbf{R}_{rr}$.

2) Compute $L_y \times 1$ filterbank output vectors $\mathbf{y}^{(\ell)}(n) := \mathbf{F}^{(\ell)H}\hat{\mathbf{T}}\mathbf{r}(n)$, for $n = 1, \ldots, M$ and $L_y \geq N$, where $\mathbf{f}_\nu^{(\ell)}$, the ν^{th} column of $\mathbf{F}^{(\ell)}$, equals $\mathbf{c}^{(\ell)}$ delayed by ν chips.
3) Estimate the normalized kurtosis of each filter output: $\hat{\kappa}_\nu^{(\ell)} := \frac{M\sum_{n=1}^{M}|y_\nu^{(\ell)}(n)|^4}{(\sum_{n=1}^{M}|y_\nu^{(\ell)}(n)|^2)^2}$.
4) Set initial equalizer $\mathbf{f}^{(\ell)}(0)$ equal to the column of $\mathbf{F}^{(\ell)}$ generating the smallest $\hat{\kappa}_\nu^{(\ell)}$.
5) Estimate ℓ^{th} user's symbols using "pre-whitened CMA": set $\mathbf{x}(n) = \hat{\mathbf{T}}\mathbf{r}(n)$, $\hat{s}_n^{(\ell)} = \mathbf{f}^{(\ell)H}(n)\mathbf{x}(n)$, and update via $\mathbf{f}^{(\ell)}(n+1) = \mathbf{f}^{(\ell)}(n) - \mu\mathbf{x}(n)\hat{s}_n^{(\ell)*}(|\hat{s}_n^{(\ell)}|^2 - 1)$.

Comment 4.1 *(Delay hypotheses):* For simplicity, 2) suggests L_y chip-spaced delay hypotheses, though better performance should result from a denser delay set (as in [1]).

Comment 4.2 *(Choice of equalization algorithm):* CMA is often considered the most widely implemented and analyzed blind equalization technique and is noted for its robustness and simplicity. Hence, it was chosen for our ME demodulation scheme. Most importantly, perhaps, CMA is known to have close ties to MMSE minimization [7] and has been shown to asymptotically achieve global MMSE solutions when $\mathcal{H}$ is full column rank and noise is absent [7]. Furthermore, CMA is known to be robust to noise and to $\mathcal{H}$ that are not full column rank [7, 9]. If speed of convergence is thought to be a problem, faster converging versions are available [15] (with higher computational cost), though pre-whitening alone greatly improves the tracking properties of CMA [14].

Comment 4.3 *(Initialization criterion):* The proposed algorithm uses kurtosis as a "measure of entropy" in the initialization stage. Recall from [8] that kurtosis-based initialization and CMA adaptation form a natural pair: $\kappa_\nu^{(\ell)} < 2$ ensures that the peak index of the asymptotic system response $\mathbf{q}^{(\ell)} := (\mathbf{f}^{(\ell)H}\hat{\mathbf{T}}\mathcal{H})^t$ equals the peak index of the initial system response. This implies that kurtosis can be directly related to acquisition probability when using codes with good cross-correlation properties.

Other initialization criteria might also work well. For example, [4] notes Gray's criterion: $\frac{M\sum_{n=1}^{M}|y_\nu(n)|^2}{(\sum_{n=1}^{M}|y_\nu(n)|)^2}$, which is easier to compute than $\hat{\kappa}_\nu^{(\ell)}$, but which simulations suggest gives comparable acquisition performance (see Figure 6).

Comment 4.4 *(Effect of noise):* In the absence of noise, pre-whitening via $\mathbf{T} = \mathbf{R}_{rr}^{-1/2}$ will result in an effective channel matrix $\mathbf{T}\mathcal{H}$ that is perfectly orthogonal. In this case, CMA regions of convergence are uniform in size and shape [16]. In the presence of noise, there is still evidence to suggest that whitening adequately reduces CMA initialization sensitivity. The pre-whitened noisy CM cost function can be written $J_{\text{cm}} = -2\|\mathbf{f}^H\mathbf{R}_{rr}^{-1/2}\mathcal{H}\|_4^4 + 3\|\mathbf{f}\|_2^4 - 2\|\mathbf{f}\|_2^2 + 1$ (for BPSK), where, remarkably, the effect of noise appears only in the first term. Note that the region of convergence sizes will be determined by $\|\mathbf{f}^H\mathbf{R}_{rr}^{-1/2}\mathcal{H}\|_4^4$ since the other terms are invariant to the orientation of $\mathbf{f}$. Even when the minimum singular value of $\mathcal{H}$ is $\approx \sigma_w$, the matrix $\mathbf{R}_{rr}^{-1/2}\mathcal{H}$ will have condition number ≈ 2, implying relatively uniform regions of convergence.

A more serious problem may occur, however, when the singular values of $\mathcal{H}$ are $< \sigma_w$, since the corresponding minima of J_{cm} may completely disappear [12]. Unfortunately, no linear pre-processing of the received signal can remedy this. In fact, this loss of weak-user minima will plague *any* demodulation scheme based on CM cost or additive modifications thereof (e.g., [10]).

Comment 4.5 *(Effect of codes/multipath):* As the codes become less orthogonal (i.e., $|r_{\ell,k}(i)|$ increases for $k \neq \ell$ or $i \neq 0$) and/or multipath becomes more severe, it can be shown that the delayed matched-filters used as initialization hypotheses (recall algorithm steps

2-4) create output streams with higher kurtosis $\kappa_\nu^{(\ell)}$. Interpreting [8] in a CDMA context, increasing $\kappa_\nu^{(\ell)}$ above 2 decreases the probability of CMA's convergence to a user with the desired code and delay ν. (Actually, this is advantageous in preventing CMA capture by an out-of-cell interferer with the desired code, since we expect out-of-cell signals to be more significantly corrupted by multipath than in-cell signals.) In addition, higher $\kappa_\nu^{(\ell)}$ increase the probability of attraction by CM saddle points, thereby potentially slowing initial convergence [9].

Comment 4.6 (Recursive estimation of $\mathbf{T}$*):* For successful acquisition in a time-varying near-far environment, it is important that $\hat{\mathbf{T}}$ tracks signal variations. Fortunately, there exist various computationally-efficient recursive strategies for updating $\hat{\mathbf{T}}$ based on a forgetting factor (e.g., using QR factorization of $\mathbf{R}_{rr}^{-1/2}$ and Given's rotations). Though the estimation of $\mathbf{T}$ is the most computationally demanding part of the algorithm, one should realize that this burden is essentially distributed over all in-cell users.

Since, when $\hat{\mathbf{T}}(n)$ is tracking, we do not want variations in $\hat{\mathbf{T}}(n)$ to adversely effect the adaptation of CMA, we recommend using CMA to adapt $\mathbf{v}^{(\ell)}(n) := \hat{\mathbf{T}}^H(n)\mathbf{f}^{(\ell)}(n)$ while operating directly on the received data: $\hat{s}_n^{(\ell)} = \mathbf{v}^{(\ell)H}(n)\mathbf{r}(n)$. The resulting filter update, $\mathbf{v}^{(\ell)}(n{+}1) = \mathbf{v}^{(\ell)}(n) - \mu\widehat{\mathbf{R}_{rr}^{-1}}(n)\mathbf{r}(n)\hat{s}_n^{(\ell)*}(|\hat{s}_n^{(\ell)}|^2{-}1)$, is Gooch's "orthogonal-CMA" [14], first proposed for synchronous CDMA applications in [11].

5 Simulation Results

In all simulations, we use Gold codes with $N{=}31$, equalizers of length 62, user delays τ_k uniformly distributed over $[0, T)$, root-raised-cosine pulse shaping with excess bandwidth 0.2, and discrete ray multipath channels $g^{(k)}(t)$. The N_g multipath rays are normally distributed in amplitude (relative to a zero-delay ray) with std. dev. σ_g and are uniformly distributed in time over the interval $[0, T_g)$. In Figures 3-4, we assume 25 in-cell users and zero out-of-cell users, while in Figures 5-6 we assume 20 in-cell users and 120 out-of-cell users, the latter with 10 dB lower average power. For in-cell users we use $N_g = 3$, $T_g = 2$, and $\sigma_g = 0.3$, while for out-of-cell users $N_g = 10$, $T_g = 10$, and $\sigma_g = 0.5$. With these parameters, the in-cell and out-of-cell ICI span approximately 10 and 18 T_c, respectively. Unless otherwise noted, we assume an in-cell log-normal user power distribution with std. dev. $\sigma_p = 5$ dB and out-of-cell with $\sigma_p = 1$ dB. The noise $\{w_i\}$ is zero-mean AWGN with SNR = 20 dB (referenced to the average in-cell user's power).

Figures 4-6 are the result of 500 Monte Carlo simulations. In Figure 4, we compare the performance of the proposed ME scheme with Madhow's multipath-specific CMOE scheme [1, Remark 3.10][4], where we assume perfect knowledge of the received signal statistics $\mathbf{R}_{rr}$ and $\kappa_\nu^{(\ell)}$. Our performance measures include both lock probability and mean-squared symbol estimation error. "Lock" occurs when final multiuser system responses $\mathbf{q}^{(\ell)}$ achieve their absolute maximum at an index corresponding to the desired user, as discussed in Comment 4.2. MSE calculations[5] are based only on receivers achieving lock. Figure 5, which includes significant out-of-cell interference, compares the performance of the ME and CMOE schemes over a range of in-cell user power deviations. Figure 6 investigates lock probability as a function of the data length M used in estimation of initialization hypotheses (recall Comment 4.3).

---

[4] We were unable to compare to [2] and [3] since both schemes require the existence of a decorrelating receiver. As discussed in Section 2.2, this is not possible in the single-sensor multi-cell environment since the anticipated number of in- plus out-of-cell users will prevent $\mathcal{H}$ from reaching full column rank.

[5] The MSE performance of the (adaptive) ME algorithm could be improved by reducing the step-size μ, whereas the performances reported for CMOE were computed in closed form.

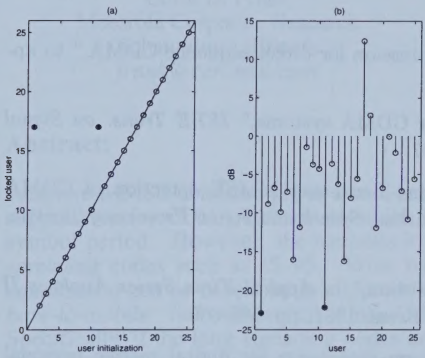


Figure 3: CMA weak-user misconvergence: (a) CMA lock from Wiener initialization, (b) user powers. (Lost users shown bold.)

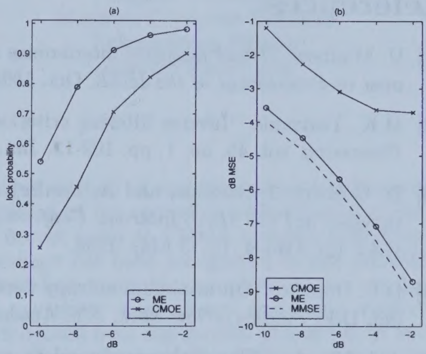


Figure 4: Lock probability and MSE performance versus desired user power (relative to 24 other equal-power users).

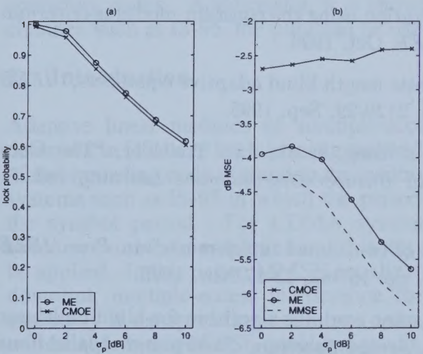


Figure 5: Lock probability and MSE performance versus in-cell (log-normal) user power deviation σ_p.

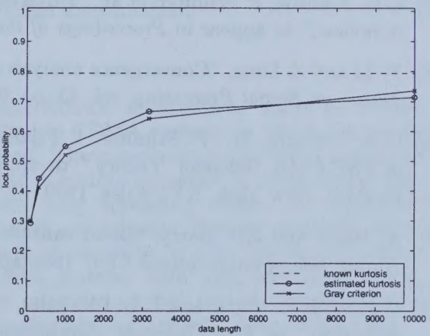


Figure 6: ME Lock probability versus data length for initialization based on data-estimated kurtosis and Gray's criterion.

6 Conclusion

We have presented a general approach to the blind acquisition/equalization of short-code CDMA based on a minimum entropy principle and, within this context, have proposed a scheme using kurtosis-based initialization of pre-whitened CMA. Simulations including multipath and significant out-of-cell interference suggest that the CMA-based method achieves higher acquisition probability and performance significantly closer to global MMSE solutions than do previously proposed schemes based on subspace-constrained optimization. Furthermore, the CMA-based approach does not require the use of multiple sensors, as do previously proposed schemes employing subspace decomposition.

Interesting questions remain concerning the relative optimality of various ME schemes and their performance in non-stationary environments.

References

[1] U. Madhow, "Blind adaptive interference suppression for direct-sequence CDMA," to appear in *Proceedings of the IEEE*, Oct. 1998.

[2] M.K. Tsatsanis, "Inverse filtering criterion for CDMA systems," *IEEE Trans. on Signal Processing*, vol. 45, no. 1, pp. 102-12, Jan. 1997.

[3] D. Gesbert, J. Sorelius, and A. Paulraj, "Blind multi-user MMSE detection of CDMA signals," in *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing* (Seattle, WA), pp. 3161-4, 12-15 May. 1998.

[4] D.L. Donoho, "On minimum entropy deconvolution," in *Applied Time Series Analysis II* (ed. D.F. Findley), New York, NY: Academic Press, 1981, pp. 565-609.

[5] A.J. Viterbi, "The orthogonal-random waveform dichotomy for digital mobile personal communication," *IEEE Personal Communications*, vol. 1, pp. 18-24, 1994.

[6] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)," *IEEE Trans. on Info. Theory*, vol. 36, no. 2, pp. 312-21, Mar. 1990.

[7] C.R. Johnson, P. Schniter, et al., "Blind equalization using the constant modulus criterion: A review," to appear in *Proceedings of the IEEE*, Oct. 1998.

[8] Y. Li and Z. Ding, "Convergence analysis of finite length blind adaptive equalizers," *IEEE Trans. on Signal Processing*, vol. 43, no. 9, pp. 2120-29, Sep. 1995.

[9] C.R. Johnson, Jr., P. Schniter, I. Fijalkow, L. Tong, et al., J.R. Treichler, "The Core of FSE-CMA Behavior Theory," to appear in *Unsupervised Adaptive Learning* (ed. S. Haykin), New York, NY: Wiley, 1999.

[10] A. Batra and J.R. Barry, "Blind cancellation of co-channel interference," in *Proc. IEEE Global Telecommunications Conf.* (Singapore), pp. 157-62, 13-17 Nov. 1995.

[11] W. Lee, B.R. Vojcic and R.L. Pickholtz, "Constant modulus algorithm for blind multiuser detection," in *Proc. Internat. Symposium on Spread Spectrum Theory and Applications* (Mainz, Germany), pp. 1262-6, 22-25 Sep. 1996.

[12] W. Chung, J.P. LeBlanc, "The local minima of fractionally-spaced CMA blind equalizer cost function in the presence of channel noise," in *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing* (Seattle, WA), pp. 3345-8, 12-15 May 1998.

[13] J.F. Cardoso, "On the performance of orthogonal source separation algorithms," in *Proc. European Signal Processing Conf.* (Edinburgh, Scotland), pp. 776-9, Sept. 1994.

[14] R. Gooch and J. Lundell, "The CM array: An adaptive beamformer for constant modulus signals," in *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing* (Tokyo, Japan), pp. 2523-6, 7-11 April 1986.

[15] B.G. Agee, "The least-squares CMA: A new technique for rapid correction of constant modulus signals," in *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing* (Tokyo, Japan), pp. 953-6, 7-11 April 1986.

[16] W. Chung and C.R. Johnson, Jr., "Characterization of the regions of convergence of CMA adaptive blind fractionally spaced equalizers," to appear in *Proc. Asilomar Conf. on Signals, Systems and Computers* (Pacific Grove, CA), 1-4 Nov. 1998.

Adaptive Interference Suppression for Direct-Sequence CDMA Systems with Long Spreading Codes

Colin D. Frank
Motorola Corporate Research
Schaumburg, Illinois
frank@ccrl.mot.com

Eugene Visotsky [†]
Motorola Corporate Research
Schaumburg, Illinois
visotsky@uiuc.edu

Abstract:

Adaptive MMSE detection [1] is a well-recognized method of multiple-access interference suppression for CDMA systems in which the period of the spreading code is equal to the symbol period. However, the methods in [1] do not apply to CDMA systems with long spreading codes such as IS-95. What has perhaps not been recognized is that MMSE *equalization* can be beneficial in suppressing downlink multiple-access interference on the *base-to-mobile link* (downlink) of CDMA systems *with long spreading codes*. Specifically, if the long spreading codes are orthogonal over the symbol period (as in IS-95), equalization can be used to eliminate intra-cell interference. For CDMA systems in which the spreading codes are not orthogonal over the symbol period, equalization can suppress non-white other-cell interference, such as that w[†] hich can be observed when only a few cells dominate the interference at the mobile receiver. In this paper, expressions are developed for the MMSE equalizer and its performance for a CDMA system with long spreading codes. Additionally, it is shown that in CDMA systems with a downlink pilot channel, such as IS-95, the pilot can be used to adapt the equalizer.

1. Introduction

Adaptive linear methods of multiple-access interference suppression, such as MMSE detection in [1], have been investigated and applied to CDMA systems in which the period of the spreading code is equal to the symbol period. These methods do not apply to CDMA systems such as IS-95 in which the period of the spreading sequence is much longer than the symbol period. For CDMA systems with long spreading codes, the *equivalent synchronous model* employed in [1] is not stationary and adaptive filtering methods cannot be applied. In this paper, MMSE equalization will be shown to be effective in suppressing downlink multiple-access interference for CDMA systems with long spreading codes, especially if the spreading codes are orthogonal over the symbol period. Furthermore, it will be shown that MMSE equalization can be implemented adaptively using a pilot code.

The nature of forward and reverse link multiple-access interference is very different, with the downlink multiple-access interference being much more structured. On the downlink, all of the signals from the desired base station (intra-cell interference) arrive at the mobile receiver via the same channel. This contrasts with the uplink in which all signals arrive at the base station receiver via a separate channel. The significance of the common channel is that, if the spreading codes are orthogonal over a symbol period, as in IS-95, all intra-cell multiple-access interference can be removed from the received signal by inverting the channel with a zero-forcing equalizer.

Other-cell interference is also much more structured on the downlink than on the uplink. Signals arriving at the mobile receiver from a base station not in communication with the mobile pass through a common channel. The channel between the interfering base station and the mobile receiver filters the interference and shapes its spectrum. If the total interference from sources other than that of the serving base is dominated by only a few of base stations, the other-cell interference may reasonably be expected to be non-white due to the spectral shaping introduced by the several channels. MMSE equalization can suppress this other-cell interference by exploiting the non-white nature of its spectrum. In the extreme in which other-cell interference dominates intra-cell interference, the MMSE equalizer takes the form of a whitened matched filter. Ultimately, however, the MMSE equalizer balances the reduction of intra-cell interference achievable by inverting the channel associated with the serving base site with the benefits of suppressing other-cell interference by exploiting its non-white spectrum.

[†] E. Visotsky performed this work while an intern with Motorola Corporate Research. He has since continued his graduate studies at the University of Illinois, Urbana, Illinois.

The organization of the rest of the paper follows. Expressions for the infinite-impulse response (IIR) MMSE equalizer and its performance are developed in Section 2. In Section 3, the expressions developed in Section 2 are used to compare the performance of the MMSE equalizer and the RAKE receiver in a two-cell system with specific propagation channels. The significance of random, rather than symbol orthogonal, long spreading codes is examined in Section 4, while in Section 5, the FIR MMSE equalizer is formulated for CDMA systems with both symbol orthogonal and random spreading codes. In Section 6, the performance of the FIR MMSE equalizer is evaluated and compared to that of the RAKE receiver.

2. MMSE Equalization for the IS-95 Forward Link

In this development, the mobile of interest is assumed to receive transmissions from two base stations. The signal received from the first base station is denoted by the sequence $\{r_i^{(1)}\}$, and is given by

$$r_i^{(1)} = \sum_{j=1}^{U_1} A_j^{(1)} \sum_{k} d_{j,k}^{(1)} \sum_{l=1}^{L_1} f_l\, c_{i-l}^{(1)}\, w_{j,i-Nk-l}\,,$$

where, for the j-th code channel, $A_j^{(1)}$ denotes the signal amplitude, $\{d_{j,k}^{(1)}\}$ denotes the data sequence, and $\{w_{j,k}\}$ denotes the Walsh code. The Walsh code $\{w_{j,k}\}$ is identically zero outside the interval $1 \leq i \leq N$, and the length of the Walsh code, N, is equal to the number of chips per code symbol. The number of occupied code channels, U_1, is less than or equal to N. The sequence $\{c_i^{(1)}\}$ denotes the long pseudo-random spreading sequence for the first base station, while the sequence $\{f_l\}$, of length L_1, denotes the impulse response of the channel between the first base station and the mobile receiver. Similarly, the signal received from the second base station is denoted by the sequence $\{r_i^{(2)}\}$, given by

$$r_i^{(2)} = \sum_{j=1}^{U_2} A_j^{(2)} \sum_{k} d_{j,k}^{(2)} \sum_{l=1}^{L_2} g_l\, c_{i-l}^{(2)}\, w_{j,i-Nk-l}\,,$$

where, for the j-th code channel, $A_j^{(2)}$ denotes the signal amplitude, $\{d_{j,k}^{(2)}\}$ denotes the data sequence, and $\{w_{j,l}\}$ denotes the Walsh code. As above, the number of occupied code channels, U_2, is less than or equal to N. The sequence $\{c_i^{(2)}\}$ denotes the long pseudo-random spreading sequence for the second base station, while the sequence $\{g_l\}$, of length L_2, denotes the impulse response of the channel between the second base station and the mobile receiver. By appropriately normalizing the amplitudes $\{A_j^{(1)}\}$ and $\{A_j^{(2)}\}$, the channels $\{f_i\}$ and $\{g_i\}$ can be assumed to have unit energy so that

$$\sum_{i=1}^{L_1} |f_i|^2 = \sum_{i=1}^{L_2} |g_i|^2 = 1\,.$$

The total signal received by the mobile receiver is given by the sequence $\{r_i\}$

$$r_i = r_i^{(1)} + r_i^{(2)} + n_i \,,$$

where $\{n_i\}$ is a white complex Gaussian noise sequence with mean sample energyN_0. As in IS-95, the first Walsh code is assumed to be used as a pilot by all base stations, so that the data sequence $b_{1,k}^{(1)} = b_{2,k}^{(2)} = 1$ for all k. The first base station is the serving base station and is assumed to transmit to the subscriber using the second Walsh code $w_{2,k}$. Since the desired signal is transmitted from the first base station, the intra-cell interference power, I_{or}, is defined as

$$I_{or} = \sum_{j=1}^{U_1} \left(A_j^{(1)}\right)^2 .$$

Interference from all sources other than the serving base station is referred to as other-cell interference, and the power of the other-cell interference, I_{oc}, is defined as

$$I_{oc} = \sum_{j=1}^{U_2} \left(A_j^{(2)}\right)^2 + N_0 \,.$$

For IS-95, the classical Rake receiver is equivalent to a matched filter followed by a despreader and Walsh correlator, as indicated in Figure 1(a). The matched filter for a signal received from the first base station has z-transform $F^*(z^{-1})$, where $F(z)$ is the z-transform of the sequence $\{f_l\}$. Let $G(z)$ denote the z-transform of the second channel $\{g_l\}$, and let the sequence $\{p_k\}$ denote the convolution $\{f_l\} * \{f^*_{-l}\}$. The signal-to-noise ratio at the output of the Walsh correlator in Figure 1(a) is then given by

$$\left(\frac{E_s}{N_t}\right)_{MF} = N\frac{E_c}{I_{or}}\left(\frac{1}{\sum p_i^2 - 1 + \left(I_{oc}/I_{or}\right) a_0}\right),$$

where

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} F\!\left(e^{j\omega}\right) F^*\!\left(e^{-j\omega}\right)\left(G\!\left(e^{j\omega}\right) G^*\!\left(e^{-j\omega}\right) + N_0\right) d\omega \,.$$

The energy-per-chip, E_c, for the channel of interest is $(A^{(1)}_2)^2\ T_c$, where T_c is the chip duration.

In Figure 1(b), the matched filter has been replaced with an equalizer. The MMSE equalizer is defined as the filter which maximizes the signal-to-noise ratio at the output of the Walsh correlator for the channel of interest $\{w_{2,k}\}$. However, MMSE equalizer can be shown to be the same for all code channels (this fact is proved for the FIR MMSE equalizer in Section (5)). This fact turns out to have three important consequences:

(i) the MMSE equalizer for the desired code *can be trained using the pilot code*;

(ii) a single MMSE equalizer can be used for all of the code channels;

(iii) the performance of all code channels is equivalent after MMSE equalization.

In order to simplify the following expressions in this section, let the AWGN component of the other-cell interference be omitted, so that N_0=0. With this restriction (for convenience), the other-cell interference can be whitened with the filter $(F(z))^{-1}$, and the MMSE equalizer has z-transform

$$C(z) = \frac{F^*(1/z)/\left(G(z)G^*(1/z)\right)}{\left[\left(F(z)F^*(1/z)\right)/\left(G(z)G^*(1/z)\right)\right] + I_{oc}/I_{or}} \,.$$

The signal-to-noise ratio at the output of the Walsh correlator is then given by

$$\left(\frac{E_s}{N_t}\right)_{MMSE} = N\frac{E_c}{I_{or}}\left(\frac{b_o}{1-b_o}\right),$$

where

$$b_o = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left(F(e^{j\omega})F^*(e^{j\omega})\right)/\left(G(e^{j\omega})G^*(e^{j\omega})\right)}{\left[\left(F(e^{j\omega})F^*(e^{j\omega})\right)/\left(G(e^{j\omega})G^*(e^{j\omega})\right)\right] + I_{oc}/I_{or}}\, d\omega\,.$$

Let Δ denote, in dB, the ratio of the signal-to-noise ratios for the MMSE and matched filter receivers , so that

$$\Delta = 10\log\left(\left(\frac{E_s}{N_t}\right)_{MMSE}\bigg/\left(\frac{E_s}{N_t}\right)_{MF}\right) = 10\ \log\left(\left(\frac{b_o}{1-b_o}\right)\left(\sum p_i^2 - 1 + (I_{oc}/I_{or})\, a_0\right)\right).$$

Note that the improvement in performance attainable with the MMSE equalizer is *independent of* E_c/I_{or}, the signal-to-interference ratio of the desired signal with respect to the serving base station. Instead, the relative performance of the two receivers depends only on the two channels, $\{f_l\}$ and $\{g_l\}$, and the ratio of the other-cell interference to intra-cell interference. This result is much different than the standard single-user equalization result in which, at low signal-to-noise ratios, the MMSE equalizer degenerates into the matched filter receiver and the benefit of equalization is diminished. The importance of this result is that it indicates that even in a heavily coded system in which the signal-to-interference ratio at the serving base might be quite low, the gains achievable through the use of MMSE equalization in the IS-95 receiver are not diminished.

3. Performance of the MMSE Equalizer in a Two-Cell Environment

The equations given above for evaluating the performance of the MMSE equalizer, the matched filter, and their relative performance are not suited for evaluating performance over fading channels. As a result, the potential benefits of MMSE equalization are evaluated by considering a few, somewhat arbitrary, channel examples for the purposes of comparison. As noted above, all that is required to evaluate the relative performance of the matched filter and MMSE equalizer are the channels $\{f_l\}$, $\{g_l\}$, and the ratio of other-cell to intra-cell interference, I_{oc}/I_{or}. The channels considered in the following examples are given in Table 1. Note that the three channels are normalized to have unit energy.

Table 1. Channels for comparison of equalizer and matched filter performance.

Channel #	channel coefficients (specified at 1 chip intervals)
1	$\{1, 1\}/\sqrt{2}$
2	$\{2, 1, 1, 1, 2\}/\sqrt{11}$
3	$\{2, 0, 1, 0, 0, 0, 0, 2, 1\}/\sqrt{10}$

In the first set of examples considered, the other-cell interference is modeled as white and Gaussian. Several scenarios can be imagined which would yield white other-cell interference. Because the spectrum of the transmitted CDMA signal is approximately flat over the bandwidth of interest, the spectrum of the interference from the second base station will be approximately white if the channel has only one path so that $G(z)=z^n$, for some n. Alternatively, other-cell interference may also be modeled as white if many base stations, rather than one or several, interfere with the mobile receiver, or if the other base stations are distant so that thermal AWGN dominates cellular interference.

Figure 2 indicates the large performance gains that can be achieved with equalization when intra-cell interference dominates the other-cell interference. The large gains achievable with equalization can be attributed almost entirely to the orthogonal spreading codes used in the IS-95 downlink. As will be discussed subsequently, these gains are not achievable using equalization if random codes, rather than orthogonal codes, are used on the downlink. Note that as other-cell interference begins to dominate intra-cell interference, the benefits of equalization diminish until the performance is essentially equivalent to that of the matched filter.

In the second set of examples, pairs of channels are taken from Table 1. One channel is selected as $\{f_l\}$, the channel from the serving base to the subscriber receiver, and a second channel is selected as $\{g_l\}$, the channel from the interfering base to the subscriber. In these examples, the AWGN component of the other-cell interference, N_0, is set equal to 0. The examples in Figure 2 indicate that the MMSE equalizer is adept at suppressing both intra-cell interference and other-cell interference. In the presence of dominant intra-cell interference, the equalizer inverts the serving base channel $F(z)$, so that $C(z)=F^{-1}(z)$, and the intra-cell interference is entirely removed. Conversely, dominant other-cell interference is suppressed by first whitening the other-cell interference, and then applying a matched filter, so that $C(z)=F^*(z^{-1})/G(z)/G^*(z^{-1})$. For the examples considered in Figure 2, in which both the serving and interfering base stations arrive at the mobile receiver via multipath channels, the gains achievable with the MMSE equalizer are greatest when one of the interference sources is dominant. The gains attainable are least significant when the two interference sources are comparable in strength.

4. Performance of MMSE Equalization with Random Codes

In the IS-95-type system described above, each base station employs the same set of N orthogonal Walsh codes of length N. The resulting spreading codes are the product of the these Walsh codes and a particular phase of a single long (length 2^{15}) spreading sequence. The N serving base station spreading codes are given by

$$\left\{c_i^{(1)} \sum_k w_{j,i-Nk},\ j \in \{1,\ldots,N\}\right\},$$

and the N spreading codes for the interfering base station are

$$\left\{c_i^{(2)} \sum_k w_{j,i-Nk},\ j \in \{1,\ldots,N\}\right\},$$

where it is understood that the N Walsh codes $\{w_{j,i}\}$ are only non-zero for $1\leq i \leq N$.

For a CDMA system with random spreading codes, each base station is assigned a set of N spreading sequences,

$$\left\{s_i^{(j)},\ j \in \{1,\ldots,N\}\right\},$$

each of which is a sequence of independent random variables taking values in the set $\{\pm 1\}$ with equal probability. Each spreading sequence is independent of both the other spreading sequences assigned to the same base station, and the sequences assigned to the other base stations in the system.

If the spreading codes in the two-cell scenario given above are replaced with random spreading codes, the MMSE equalizer is unchanged; that is,

$$C(z) = \frac{F^*(1/z)/(G(z)G^*(1/z))}{[(F(z)F^*(1/z))/(G(z)G^*(1/z))] + I_{oc}/I_{or}} .$$

However, the performance of the MMSE equalizer for the two systems is not the same. Let SNR_O and SNR_R denote the MMSE signal-to-noise ratio for orthogonal codes and random codes, respectively. For the two-cell scenario described above, the two signal-to-noise ratios can be shown to satisfy the relation

$$SNR_R = \frac{1}{(SNR_O)^{-1} + ((E_c/I_{or})^{-1} - 1)} .$$

The second term in the denominator essentially accounts for the intra-cell interference. Since $E_c/I_{or} < 1$, the second term in the denominator is non-negative and $SNR_R < SNR_O$.

With respect to the MMSE equalizer, the primary significance of the use of random codes rather than orthogonal codes within a base site is that the equalizer can no longer effectively suppress intra-cell interference by inverting the channel. In contrast, the effectiveness of the equalizer in suppressing interference from a second base station does not depend on whether the codes within that base station are orthogonal to each other or not. Rather, the ability of the MMSE equalizer to suppress other-cell interference depends only on the second order statistics of the other-cell interference which, in turn, depends only on the channel $\{g_l\}$ and not on whether the spreading codes are random or orthogonal. Thus, in summary, the MMSE equalizer can effectively suppress other-cell interference in a CDMA system with random codes, but the use of random codes reduces the ability of the equalizer to suppress intra-cell interference.

5. Finite Impulse Response Implementation of the MMSE Equalizer

In this section, the finite-impulse response (FIR) MMSE equalizer and its corresponding performance are developed for the two-cell scenario presented above. As above, it is assumed that the desired signal is transmitted on the second code channel $\{w_{2,k}\}$. Let **R** denote a vector of length M with components

$$R_l = \sum_{i=1}^{N} r_{l+i}\, c_i^{(1)}\, w_{2,i}, \quad l \in \{1,\ldots, M\} .$$

The MMSE equalizer is the vector **q** which minimizes the mean-square error

$$E(|\mathbf{R}^* \mathbf{p} - 1|^2) .$$

over all vectors **p** of length M. The equalizer **q** can be expressed as

$$\mathbf{q} = (E(\mathbf{R}\mathbf{R}^*))^{-1} E(\mathbf{R})$$

$$= \Gamma^{-1} \mu ,$$

where the length-M vector μ and the $M \times M$ matrix Γ are, respectively, the mean and covariance of **R**.

The mean vector μ is given by

$$\mu_l = E(R_l) = N A_2^{(1)} f_l .$$

If the spreading codes are orthogonal, the elements of the covariance matrix Γ are given by

$$\Gamma_{l,m} = E\left(R_l R_m^*\right) = \left(N\sum_{j=1}^{U_1}\left(A_j^{(1)}\right)^2 \sum_{k\neq l} f_k f_{m-l+k}^* \right) + N^2\left(A_2^{(1)}\right)^2 f_l f_m^* + N\varphi(m-l),$$

where φ(·) is the covariance of the other-cell interference. For the two-cell scenario given above, the covariance is given by

$$\varphi(m) = \left(\sum_{j=1}^{U_2}\left(A_k^{(2)}\right)^2\right)\left(\sum_l g_l g_{l+m}^*\right) + N_0 .$$

Alternatively, if the spreading codes are random, the elements of the covariance matrix are given by

$$\Gamma_{l,m} = \left(N\sum_{j=1}^{U_1}\left(A_j^{(1)}\right)^2 \sum_{k\neq l} f_k f_{m-l+k}^* \right) + N^2\left(A_2^{(1)}\right)^2 f_l f_m^* + N\varphi(m-l)$$
$$+ N\left(\sum_{j\neq 2}\left(A_j^{(1)}\right)^2\right) f_l f_m^* .$$

In the discussion of the IIR MMSE equalizer presented above, the following facts were used (without proof):

Claim 1: In a CDMA system with orthogonal spreading codes, the MMSE equalizer is the same for all spreading codes (within a multiplicative constant).

Claim 2: The MMSE equalizer is the same for a CDMA system using orthogonal codes as for a CDMA system using random spreading codes (within a multiplicative constant).

As noted previously, Claim 1 is particularly important because it implies that the pilot channel can be used to train the MMSE equalizer for any channel of interest. A second consequence of Claim 1 is that all of the code channels can be demodulated with a single equalizer, and this is an important property for a multi-code receiver.

Claims 1 and 2 will now be proven for the FIR MMSE equalizer. To begin, note that the mean vector μ is scaled by the amplitude of the desired code channel. However, in all cases the mean vector is of the form

$$\mu = \alpha \mathbf{f} ,$$

where **f** is a vector of the channel coefficients $\{f_l\}$, and the scalar α depends on the code channel of interest. Next, refer to the expressions for the covariance matrix Γ and note that, regardless of the particular channel or whether the codes are orthogonal or random, the covariance matrix is always of the form

$$\Gamma = \Lambda + \beta \mathbf{f}\mathbf{f}^* ,$$

for some scalar β, where the matrix Λ is defined as

$$\Lambda_{l,m} = \left(N\sum_{j=1}^{U_1}\left(A_j^{(1)}\right)^2 \sum_{k\neq l} f_k f_{m-l+k}^* \right) + N\varphi(m-l) .$$

The scalar β depends both on the code channel of interest and on whether the codes are orthogonal or random.

For general α and β, the MMSE equalizer is given by

$$\begin{aligned}\mathbf{q} &= \Gamma^{-1}\,\boldsymbol{\mu} \\ &= \left(\Lambda + \beta\mathbf{f}\right)^{-1}\alpha\mathbf{f}\,.\end{aligned}$$

Using the matrix inversion lemma, the MMSE equalizer **q** can be expressed as

$$\begin{aligned}\mathbf{q} &= \left(\Lambda + \beta\mathbf{f}\,\mathbf{f}^*\right)^{-1}\alpha\mathbf{f} \\ &= \alpha\left(\Lambda^{-1}\,\mathbf{f} + \frac{\beta}{1+\lambda}\Lambda^{-1}\,\mathbf{f}\,\mathbf{f}^*\,\Lambda^{-1}\,\mathbf{f}\right) \\ &= \alpha\left(1 + \frac{\beta}{1+\lambda}\mathbf{f}^*\,\Lambda^{-1}\,\mathbf{f}\right)\Lambda^{-1}\,\mathbf{f} \\ &= \kappa\,\Lambda^{-1}\,\mathbf{f}\,,\end{aligned}$$

for appropriate constants λ and κ. Thus, except for a multiplicative constant κ, the MMSE equalizer is independent of either the amplitude of the code channel of interest or whether the spreading codes are orthogonal or random. Thus, the pilot code can be used to train the MMSE equalizer, and one MMSE equalizer can be used for all of the code channels of interest.

6. Performance of the FIR MMSE Equalizer

The signal-to-noise ratio following the FIR MMSE equalizer and correlation with the desired code channel is given by

$$\left(\frac{E_s}{N_t}\right)_{\mathrm{FIR,MMSE}} = \frac{\boldsymbol{\mu}^*\,\Gamma^{-1}\,\boldsymbol{\mu}}{1-\boldsymbol{\mu}^*\,\Gamma^{-1}\,\boldsymbol{\mu}}\,.$$

Using the matrix inversion lemma, the signal-to-noise ratio can also be expressed as

$$\left(\frac{E_s}{N_t}\right)_{\mathrm{FIR,MMSE}} = N\frac{E_c}{I_{or}}\mathbf{f}^*\Omega^{-1}\,\mathbf{f}\,,$$

where the matrix Ω is defined to have elements given by

$$\Omega_{l,\,m} = \sum_{k\neq l} f_k\, f^*_{m-l+k} + \frac{I_{oc}}{I_{or}}\varphi(m-l)\,.$$

The signal-to-noise ratio following the matched filter and correlation with the desired code channel is

$$\left(\frac{E_s}{N_t}\right)_{\mathrm{MF}} = \frac{\left(\boldsymbol{\mu}^*\boldsymbol{\mu}\right)^2}{\boldsymbol{\mu}^*\Gamma\,\boldsymbol{\mu} - \left(\boldsymbol{\mu}^*\boldsymbol{\mu}\right)^2}\,.$$

This expression can also be written as

$$\left(\frac{E_s}{N_t}\right)_{\mathrm{MF}} = N\frac{E_c}{I_{or}}\frac{\left(\mathbf{f}^*\mathbf{f}\right)^2}{\mathbf{f}^*\Omega\,\mathbf{f}}\,.$$

Using the above expressions for the signal-to-noise ratio of the finite implementations of the MMSE equalizer and matched filter, respectively, the ratio Δ denote of their relative performance is given by

$$\Delta = 10\log\left(\left(\frac{E_s}{N_t}\right)_{MMSE} \Big/ \left(\frac{E_s}{N_t}\right)_{MF}\right) = 10\ \log\left(\frac{(\mathbf{f}^*\Omega^{-1}\mathbf{f})(\mathbf{f}^*\Omega\mathbf{f})}{(\mathbf{f}^*\mathbf{f})^2}\right).$$

From the definitions of Ω and **f**, it is again apparent that the relative performance Δ of the MMSE equalizer and the matched filter does not depend on E_c/I_{or}. Rather, as for the IIR implementation of the equalizer, the relative performance of the FIR MMSE equalizer and the matched filter depends only on the channel between the serving base and the subscriber, the covariance of the other-cell interference, and the ratio, I_{oc}/I_{or}, of the power of the other cell interference to that of the intra-cell interference. Thus, even in a heavily coded system in which the signal-to-interference ratio at the serving base might be quite low, the gains achievable through the use of MMSE equalization in the IS-95 receiver are not diminished.

Acknowledgments:

The authors would like to Motorola Corporate Research for supporting this work and allowing its publication. The authors would also like to gratefully acknowledge several helpful discussions with Upamanyu Madhow.

References:

[1] U. Madhow and M. Honig, "MMSE interference suppression for direct-sequence spread spectrum CDMA," *IEEE Trans. on Comm.*, vol. 42, pp. 3178-3188, Dec. 1994.

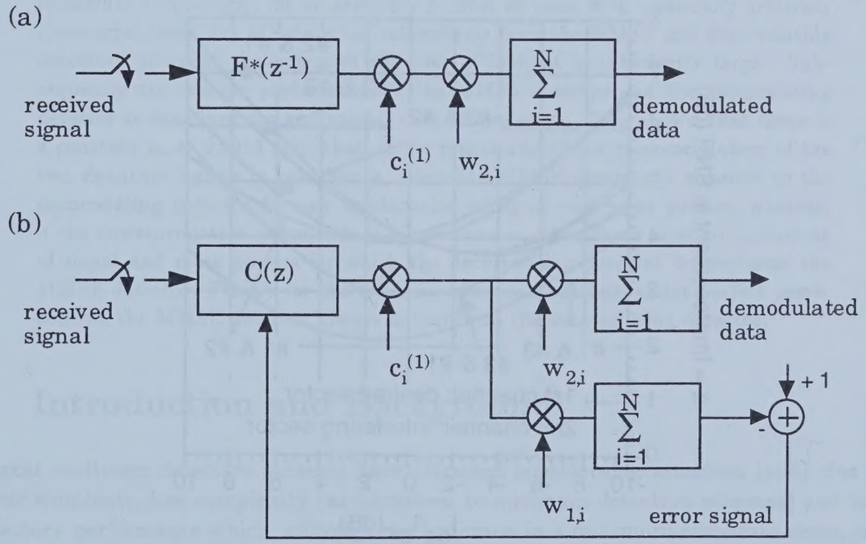


Figure 1. (a) Matched filter (RAKE) receiver for IS-95 forward link.
(b) MMSE equalization of the IS-95 forward link. The pilot channel is used to train the equalizer.

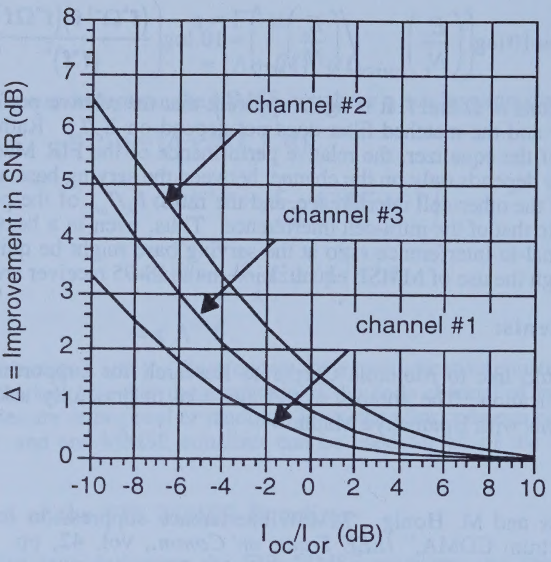


Figure 2. The relative performance of the MMSE equalizer and the matched filter as a function of I_{oc}/I_{or}. The other-cell interference is modeled as AWGN.

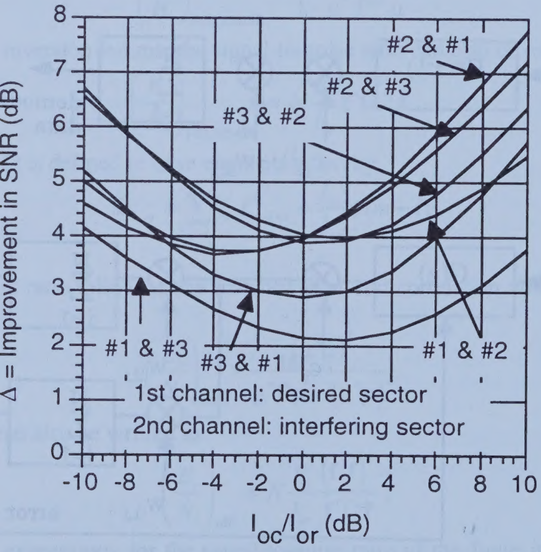


Figure 3. The relative performance of the MMSE equalizer and the matched filter as a function of I_{oc}/I_{or}. The other-cell interference is modeled as AWGN filtered by a multipath channel.

On the Relative Error Probabilities of Linear Multiuser Detectors*

George V. Moustakides
Dept. of Computer Engineering and Informatics
University of Patras
Patras, GREECE
moustaki@cti.gr

H. Vincent Poor
Dept. of Electrical Engineering
Princeton University
Princeton, NJ 08544 USA
poor@princeton.edu

Abstract

The relative error-probability performance of three linear multiuser detectors - the minimum mean square error (MMSE) detector, the decorrelator, and the conventional matched filter (MF) detector - is investigated under nonorthogonal signaling and additive white Gaussian noise conditions. It is first shown that, somewhat surprisingly, for an arbitrary number of users with essentially arbitrary crosscorrelations, the MF detector outperforms both the MMSE and decorrelating detectors provided the power of the user of interest is sufficiently large. Subsequently the relative performance of the MMSE detector and the decorrelating detector is examined for the special case of two users. It is shown that there is a constant $\rho_\star \approx 0.9918$ such that, if the magnitude of the crosscorrelation of the two signature signals is less than $\rho_\star$, then the MMSE detector is superior to the decorrelating detector for any combination of signal and noise powers; whereas, if the crosscorrelation magnitude is larger than $\rho_\star$, then there exist combinations of signal and noise powers for which the decorrelating detector outperforms the MMSE detector. Finally for the two-user case we show that, under perfect power control, the MMSE detector always outperforms the decorrelating detector.

I Introduction and Background

Linear multiuser detection schemes have attracted considerable attention lately due to their simplicity, low complexity (as compared to optimum detection schemes) and satisfactory performance which, although not optimum in a minimum-error-rate sense, can nevertheless satisfy a number of alternative asymptotic optimization criteria such as high efficiency or near-far resistance.

The matched filter (MF) detector is, of course, the simplest linear detector. Since this detector neglects the presence of interfering users, its performance can be extremely poor in the presence of severe multiple access interference (MAI). Two key linear multiuser

---

*This work was supported in part by the National Science Foundation and in part by a Collaborative Research Grant from the NATO International Scientific Exchange Programme.

detectors are known to combat MAI very efficiently (see, for example, [3]). Specifically the *decorrelating* detector (or *decorrelator*) defeats MAI by selecting a proper linear filter to eliminate it, while the *minimum mean square error* (MMSE) detector achieves robustness against MAI by selecting the linear filter that minimizes the mean square value of the output MAI plus noise.

The MMSE detector has been the center of recent attention due to its noticeable feature of being practically implementable through blind adaptive schemes; that is, through schemes that use only the signature waveform of one user of interest and do not require training sequences or knowledge of signatures of interferers [1]. In [2] one can find a detailed analysis of this detector's performance under various conditions related to multiuser applications, along with efficient approximations of the corresponding error rates.

Several analytical and numerical results have suggested the conjecture (stated in [2]) that the MMSE detector outperforms the decorrelator for any combination of signal and noise powers. Furthermore a similar (if not stronger) feeling seems to be shared when the MMSE detector is compared against the conventional MF detector. It is the aim of this work to prove that both of these conjectures are in fact false. Specifically, for the comparison of the MMSE detector with the decorrelator we show that for the two-user case and for large enough crosscorrelation of the signature signals, it is possible to find noise and signal powers for which the MMSE detector is inferior to the decorrelator. We also show that it is not possible to find such cases under the condition of perfect power control. For the comparison of the MMSE detector against the MF detector we will show that, for the general K-user case with essentially arbitrary crosscorrelations, the MF detector outperforms the MMSE detector provided that the power of the user of interest is sufficiently large. We also show a similar result for the MF detector versus the decorrelator.

Let us now define the problem of interest more specifically. Consider a K-user binary communication system with corresponding normalized modulation waveforms $s_1, \ldots, s_K$, and signaling antipodally through an additive white Gaussian noise channel. If we limit ourselves to the synchronous signal case then the length-K vector $\boldsymbol{y}$ whose l-th component is the output of a filter matched to s_l is a sufficient statistic for the problem of detecting the transmitted symbols. The vector $\boldsymbol{y}$ can be written as [3, Page 56]

$$\boldsymbol{y} = \boldsymbol{R}\boldsymbol{A}\tilde{\boldsymbol{b}} + \sigma\boldsymbol{n} \tag{1}$$

where $\boldsymbol{R}$ denotes the normalized crosscorrelation matrix of the signals $s_1, \ldots, s_K$; $\boldsymbol{A}$ is the diagonal matrix $\text{diag}\{A_1, \ldots, A_K\}$ with A_l the received amplitude of user-l; $\tilde{\boldsymbol{b}} = [b_1 \cdots b_K]^T$ is a vector whose l-th component is the symbol $b_l \in \{\pm 1\}$ of user-l; $\boldsymbol{n}$ is a $\mathcal{N}(0, \boldsymbol{R})$ random vector independent of $\tilde{\boldsymbol{b}}$, and finally σ^2 is the intensity of the additive channel noise.

As noted above we wish to examine the relative performance of the MMSE detector, the decorrelator and the MF detector. Without loss of generality throughout our analysis we will consider User # 1 to be the user of interest. The corresponding error probabilities are then given by the following formulas [3, Pages 113, 249, 300]:

$$P_{MMSE} = 2^{1-K} \sum_{b_2,\ldots,b_K \in \{-1,1\}^{K-1}} Q\left(\frac{\boldsymbol{e}_1^T(\boldsymbol{R}+\boldsymbol{X}^{-2})^{-1}\boldsymbol{R}\boldsymbol{X}\boldsymbol{b}}{\sqrt{\boldsymbol{e}_1^T(\boldsymbol{R}+\boldsymbol{X}^{-2})^{-1}\boldsymbol{R}(\boldsymbol{R}+\boldsymbol{X}^{-2})^{-1}\boldsymbol{e}_1}}\right) \tag{2}$$

$$P_D = Q\left(x_1/\sqrt{e_1^T R^{-1} e_1}\right) \tag{3}$$

$$P_{MF} = 2^{1-K} \sum_{b_2,\ldots,b_K \in \{-1,1\}^{K-1}} Q\left(e_1^T R X b\right) \tag{4}$$

where $X = \mathrm{diag}\{x_1, \ldots, x_K\}$ with $x_l = A_l/\sigma$; $b = [1\ b_2 \cdots b_K]^T$; $e_1 = [1\,0 \cdots 0]^T$ and finally $Q(\cdot)$ denotes the complementary unit cumulative Gaussian distribution. In the following sections, we compare these three expressions under various signaling conditions.

II The Matched Filter versus the MMSE and Decorrelating Detectors

In this section our goal is to show that *the MF outperforms both the MMSE and decorrelating detectors provided the user of interest is sufficiently strong in power.* In particular, we prove the following result.

Proposition 1: Fix $\boldsymbol{R}, x_2, \ldots, x_K$*, and assume that* $\boldsymbol{R}$ *is positive definite. If at least one interfering user has a signature waveform that is nonorthogonal to the signature waveform of User # 1, then there exists sufficiently large* x_1 *for which the MF detector outperforms the MMSE detector and the decorrelator.*

Proof: The error probabilities for the detectors of interest are given in (2) - (4). It is convenient, however, to rewrite these expressions in order to reveal the linear dependency on x_1 of the arguments of the Q-functions. Since $X^{-2} = x_1^{-2} e_1 e_1^T + S$ with $S = \mathrm{diag}\{0, x_2^{-2}, \ldots, x_K^{-2}\}$ the matrix inversion lemma yields

$$(\boldsymbol{R} + X^{-2})^{-1} e_1 = \frac{(\boldsymbol{R} + S)^{-1} e_1}{1 + x_1^{-2} e_1^T (\boldsymbol{R} + S)^{-1} e_1}. \tag{5}$$

Substituting this into the expression (2) for the error probability for the MMSE detector and noting that

$$e_1^T (\boldsymbol{R} + S)^{-1} \boldsymbol{R} e_1 = e_1^T (\boldsymbol{R} + S)^{-1} (\boldsymbol{R} + S - S) e_1 = 1 \tag{6}$$

(because $S e_1 = 0$), we obtain the following alternative expression

$$P_{MMSE} = 2^{1-K} \sum_{b_2,\ldots,b_K \in \{-1,1\}^{K-1}} Q\left(\frac{x_1 + e_1^T (\boldsymbol{R} + S)^{-1} \boldsymbol{R} \tilde{X} b}{\sqrt{e_1^T (\boldsymbol{R} + S)^{-1} \boldsymbol{R} (\boldsymbol{R} + S)^{-1} e_1}}\right), \tag{7}$$

where $\tilde{X} = \mathrm{diag}\{0, x_2, \ldots, x_K\}$. It is also convenient to display explicitly the dependence of the MF detector's performance on x_1; namely,

$$P_{MF} = 2^{1-K} \sum_{b_2,\ldots,b_K \in \{-1,1\}^{K-1}} Q\left(x_1 + e_1^T \boldsymbol{R} \tilde{X} b\right). \tag{8}$$

It should be noted that none of the quantities in bold face in (7) and (8) depends on x_1.

Using (6), the fact that $e_1^T \boldsymbol{R} e_1 = 1$, and the Schwarz inequality, we can show that

$$\frac{1}{\sqrt{e_1^T (\boldsymbol{R} + S)^{-1} \boldsymbol{R} (\boldsymbol{R} + S)^{-1} e_1}} \leq 1, \tag{9}$$

with equality iff

$$R^{1/2}e_1 = \alpha R^{1/2}(R+S)^{-1}e_1 \tag{10}$$

for some scalar α. It is easy to verify that (10) holds iff $\rho_{l1} = 0,\ l = 2, \ldots, K$, where ρ_{l1} is the l-th component of the first column of R. In other words we have equality in (9) iff simultaneously all interfering users have signature waveforms that are orthogonal to the signature waveform of User # 1. If at least one interfering user does not satisfy this constraint, then the inequality in (9) is strict.

Now, from (7) and (8), it follows immediately that, with (9) strict and for all sufficiently large x_1, we will have $P_{MMSE} > P_{MF}$. This is because each of these error probabilities is dominated in the tails (of large values of x_1) by the term involving the Q-function with the smallest argument. If (9) is strict, then for sufficiently large x_1 the smallest such argument of P_{MMSE} will be smaller than the smallest such argument in P_{MF}.

Similarly we can show the corresponding result for the decorrelator. Notice that using again the Schwarz inequality we can show

$$1 = e_1^T R^{-1/2} R^{1/2} e_1 \leq \sqrt{e_1^T R^{-1} e_1}\sqrt{e_1^T R e_1} = \sqrt{e_1^T R^{-1} e_1} \tag{11}$$

with equality iff $\alpha e_1 = Re_1$ for some scalar α or equivalently iff $\rho_{1l} = 0,\ l = 2 \ldots, K$. The rest of the proof goes exactly as in the previous case. ■

For the case where the signatures of the interfering users are orthogonal to the signature of User # 1 we know that all three schemes have the same performance $P_{MMSE}(x_1) = P_D(x_1) = P_{MF}(x_1) = Q(x_1)$, which is of course the single-user performance.

III The MMSE Detector versus the Decorrelating Detector

We now turn, as in [2], to a comparison of the MMSE detector and the decorrelating detector. Here, we restrict attention to the two-user case ($K = 2$). Proceeding along the same lines of [2, Proposition 5.2] we will show the following result. If ρ denotes the crosscorrelation of the (normalized) signature waveforms of the two users, then the MMSE detector outperforms the decorrelator provided that $|\rho|$ is smaller than some upper limit $\rho_\star < 1$. The significant new information brought by our result as compared to [2] (apart a slight improvement on the upper bound for $|\rho|$) is the fact that the proposed upper bound is in fact a tight one. By this we mean that *if* $|\rho| > \rho_\star$ *then there is a combination of noise and signal powers for which the decorrelator outperforms the MMSE detector.*

It should be noted that, due to our last comment, we will be able to show that the conjecture stated in [2] (that the MMSE detector is always better than the decorrelator) is in fact false. At the end of this section we will also give an explicit counterexample to this conjecture.

To study this problem, it is convenient to rewrite the error probabilities for the two detectors using a slightly different notation. In particular, the error probabilities of interest can be written as follows

$$P_{MMSE} = \frac{1}{2}Q(ax_1 + b) + \frac{1}{2}Q(ax_1 - b) \tag{12}$$

and

$$P_D = Q(cx_1) \tag{13}$$

where

$$a = \frac{1 - \rho^2 + \phi_2}{\sqrt{(1 + 2\phi_2)(1 - \rho^2) + \phi_2^2}}\ ,\ b = a\frac{\rho\sqrt{\phi_2}}{1 - \rho^2 + \phi_2},\ c = \sqrt{1 - \rho^2}\ ,\ \text{and}\ \phi_2 = x_2^{-2}. \tag{14}$$

Notice that P_{MMSE} and P_D are symmetric in the correlation factor ρ so, without loss of generality, we will assume that $\rho \geq 0$. Let ϕ_2 and ρ be fixed. Then the parameters a, b, c also become fixed since they depend only on ϕ_2 and ρ and not on x_1. The two error probabilities can thus be written as functions of x_1, and we denote their difference by

$$J(x_1) = P_{MMSE}(x_1) - P_D(x_1). \tag{15}$$

In the following subsection we analyze this function in detail.

A. Analysis of $J(x_1)$

We first note that $J(0) = J(\infty) = 0$. Consider now the derivative of $J(x_1)$, which takes the form

$$J'(x_1) = \frac{1}{\sqrt{2\pi}}\left\{ce^{-c^2x_1^2/2} - \frac{1}{2}a\left[e^{-(ax_1+b)^2/2} + e^{-(ax_1-b)^2/2}\right]\right\}. \tag{16}$$

Before studying the sign of this derivative let us compare the two parameters a and c. From (14) it is easy to verify that a is strictly decreasing in ϕ_2. Since for $\phi_2 = 0$ the two quantities a and c become equal, we conclude that for any $\phi_2 > 0$ we have $a > c$.

To analyze the sign of the derivative $J'(x_1)$ it suffices to analyze the sign of the expression

$$\frac{\sqrt{2\pi}}{a}e^{c^2x^2/2}J'(x_1) = d - F(x_1) \tag{17}$$

where

$$F(x_1) = \frac{1}{2}\left\{e^{-(c_1x_1+c_2)^2/2} + e^{-(c_1x_1-c_2)^2/2}\right\}, \tag{18}$$

$$d = \frac{c}{a}e^{-b^2c^2/2(a^2-c^2)}\ ,\ c_1 = \sqrt{a^2 - c^2}\ ,\ \text{and}\ c_2 = ab/\sqrt{a^2 - c^2}. \tag{19}$$

A typical form of the function $F(x_1)$ is depicted in Fig. 1. Notice that the number of sign alternations of the expression in (17) for $x_1 \geq 0$ depends on the number of positive real roots of the equation $d = F(x_1)$. This equation has at least one positive root (since the relation $J(0) = J(\infty) = 0$ guarantees one extremum for $J(x_1)$) and at most two (see Fig. 1). We can now prove the following lemma.

Lemma 1: Suppose ϕ_2 and ρ are fixed. Then the MMSE detector is better than the decorrelator for every value of x_1 iff $d \leq F(0)$.

Proof: If $d \leq F(0)$ then we note from Fig. 1 that the equation $d = F(x_1)$ has only one positive root (say $x_\star$). This means that $J(x_1)$ is decreasing in x_1 for $0 \leq x_1 \leq x_\star$ and increasing for $x_1 \geq x_\star$. Since $J(0) = J(\infty) = 0$ this suggests that $J(x_1) \leq 0$; that is, that the MMSE detector is superior to the decorrelator for all values of x_1.

Conversely, if $d > F(0)$ then from Fig. 1 we observe that the MMSE detector is inferior to the decorrelator for values of x_1 sufficiently close to zero (since $J'(0) > 0$). ■

Let us now substitute all quantities entering in the inequality $d \leq F(0)$ in terms of the two parameters ρ and ϕ_2. We then obtain the following equivalent relation

$$G(\phi_2, \rho) = \frac{1 - \rho^2 + \phi_2}{\sqrt{1-\rho^2}\sqrt{(1+2\phi_2)(1-\rho^2)+\phi_2^2}}\, e^{\frac{-\rho^2\phi_2}{2(1+2\phi_2)(1-\rho^2)+\phi_2^2}} \geq 1 \qquad (20)$$

which due to Lemma 1 is necessary and sufficient for guaranteeing superior performance for the MMSE detector over the decorrelator for every value of x_1.

From the proof of Lemma 1 it is clear that if (20) is not satisfied for some combination of ϕ_2 and ρ, then we can then find small enough x_1 such that $P_{MMSE} > P_D$.

In order to proceed with our goal, namely to find all values of ρ for which the MMSE detector is better than the decorrelator for any value of x_1 and ϕ_2, we need to study the inequality in (20). This is the subject of the next subsection.

B. Analysis of $G(\phi_2, \rho)$

Let us fix ρ and consider $G(\phi_2, \rho)$ as a function of ϕ_2. We can then verify that the sign of its partial derivative with respect to ϕ_2 is the same as the sign of the following third order polynomial

$$U(\phi_2, \rho) = (1 - \rho^2 + \phi_2)^3 - \rho^2(1-\rho^2)(1-\rho^2+\phi_2) + 2\rho^2(1-\rho^2)^2. \qquad (21)$$

Since the constant term of this polynomial is positive, the sign of the product of its three roots is negative. This in turn suggests that the polynomial has one real negative root and, depending on ρ, either two additional complex conjugate roots or two additional real roots that are of the same sign.

Using standard results concerning roots of third-order polynomials one can show that when $\rho > \rho_m = \sqrt{27/28} \approx 0.982$ the polynomial has three distinct real roots (one of which is negative), whereas for $0 < \rho \leq \rho_m$, the polynomial has one negative real root and two complex conjugate roots.

Consider the case $\rho > \rho_m$; as noted above, the polynomial $U(\phi_2, \rho)$ has three distinct real roots in this case, one of which is negative. We now need to identify the common sign of the two other real roots. If we substitute $\phi_2 = 2(1-\rho^2)$ then $U(2(1-\rho^2), \rho) < 0$ meaning that there is at least one root in the interval $(2(1-\rho^2)\ , \infty)$, and consequently the remaining two roots are both positive. From the above we have the following remarks concerning the function $G(\phi_2, \rho)$.

Remark 1: If $\rho \leq \rho_m$, then $U(\phi_2, \rho)$ is positive for $\phi_2 \geq 0$, and consequently $G(\phi_2, \rho)$ is strictly increasing in ϕ_2 and larger than unity (since $G(0, \rho) = 1$). For this case the inequality in (20) is clearly satisfied for every $\phi_2 \geq 0$.

Remark 2: If $\rho > \rho_m$ then $G(\phi_2, \rho)$, as a function of ϕ_2 and for $\phi_2 \geq 0$, presents two local extrema at the two positive real roots of $U(\phi_2, \rho)$, the first one being a local maximum and the second a local minimum. Since the first extremum is a local maximum this means that its value is larger than unity (because $G(0, \rho) = 1$). The local minimum however can be either larger or smaller than unity depending on ρ. If for some ρ the corresponding local minimum of $G(\phi_2, \rho)$ goes below unity then inequality (20) is clearly not satisfied for every $\phi_2 \geq 0$. If on the other hand the local minimum is greater or equal than unity then (20) is satisfied for all $\phi_2 \geq 0$.

The variable behavior of $G(\phi_2, \rho)$ is depicted in Fig. 2 where we plot this function with respect to ϕ_2 and for characteristics values of the parameter ρ. We are now in a position to find all values of the crosscorrelation ρ for which the MMSE detector outperforms the decorrelator for all possible noise and signal powers.

Proposition 2: There exists a $\rho_\star$ with $\sqrt{27/28} < \rho_\star < 1$ such that if $0 \leq \rho \leq \rho_\star$ then the MMSE detector outperforms the decorrelator for any value of x_1 and ϕ_2; furthermore if $1 \geq \rho > \rho_\star$ then there exist values for x_1 and ϕ_2 such that the MMSE detector is inferior to the decorrelator.

Proof: Due to Remark 1 we conclude that for any $0 \leq \rho \leq \rho_m$ we have that $G(\phi_2, \rho) \geq 1$ for all $\phi_2 \geq 0$. This of course implies that for any $\rho \in [0, \rho_m]$ the MMSE detector outperforms the decorrelator for any noise and signal powers. With the help of Remark 2 we will be able to slightly improve this result and propose an upper bound on ρ that is in fact tight.

Consider the case $1 \geq \rho > \rho_m$ where Remark 2 applies. If we plot the local minimum of $G(\phi_2, \rho)$ as a function of ρ then from Fig. 3 we have that this function is decreasing in ρ. Furthermore there exists a value $\rho_\star$ for which the corresponding local minimum is exactly equal to unity. Due to Remark 2 and the monotonicity of the local minimum as a function of ρ we then conclude that for $\rho_\star \geq \rho > \rho_m$ we have $G(\phi_2, \rho) \geq 1$ for all values of ϕ_2. On the other hand if $1 \geq \rho > \rho_\star$, there exist values for ϕ_2 where the inequality $G(\phi_2, \rho) \geq 1$ is false (see Fig. 2).

Combining the two intervals we conclude that for any $0 \leq \rho \leq \rho_\star$ the inequality in (20) is true for all values of ϕ_2 therefore the MMSE detector outperforms the decorrelator for any noise and signal powers. The upper limit $\rho_\star$ can be computed numerically; the value we obtain is $\rho_\star = 0.991765239964$.

From the above discussion we can also deduce that the proposed upper limit $\rho_\star$ is tight since for any $1 \geq \rho > \rho_\star$ we can find values for the parameter ϕ_2 such that (20) is false, meaning that there exists sufficiently small x_1 such that $J(x_1) > 0$. ■

From the analysis it is also apparent that if we had used x_2 instead of ϕ_2 this would have resulted in a function $G(x_2, \rho)$ with local extrema appearing at (very) large values of x_2 thus complicating the presentation.

C. Numerical (Counter) Example

A counter example for the conjecture in [2] is the following. Let $\rho = 0.995 > \rho_\star$, from Fig. 2 we then observe that if we select for example $\phi_2 = 0.05$ then inequality (20) is false, indeed we have $G(0.05, 0.995) \approx 0.824 < 1$. According to our analysis there exists small enough x_1 such that the MMSE detector is inferior to the decorrelator. If we select x_1 even as large as $x_1 = 1$ this yields $P_{MMSE} = 0.4634 > P_D = 0.4602$.

D. Power Control

In this subsection we are going to consider the special case $A_1 = A_2 = A$ or $x_1 = x_2 = x$ ($\phi_2 = x^{-2}$) for the two-user case. The relative performance of the two detectors for this practically interesting case is analyzed in the next proposition.

Proposition 3: For the two-user case and under perfect power control the MMSE detector always outperforms the decorrelator.

Proof: Notice from (12) and (13) that for $\rho = 1$ and for $x_1 = x$, $\phi_2 = x^{-2}$ we have $P_{MMSE}(x) = P_D(x)$. Let us therefore consider the case $\rho < 1$. We can now apply the following change of variables $x = \xi/\sqrt{1-\rho^2}$ after which the corresponding error probabilities then become

$$P_{MMSE}(\xi,\rho) = \frac{1}{2}\sum_{b_2\in\{\pm 1\}} Q\left(\xi\frac{\xi^2+1+b_2\rho}{\sqrt{\xi^4+(1-\rho^2)(2\xi^2+1)}}\right), \tag{22}$$

and $P_D(\xi) = Q(\xi)$.

Our intention next is to show that $P_{MMSE}(\xi,\rho)$, for fixed ξ, is a decreasing function of ρ. This property will be sufficient to prove our proposition because we can then write $P_{MMSE}(\xi,\rho) \leq P_{MMSE}(\xi,0) = Q(\xi) = P_D(\xi)$.

To show that the partial derivative of $P_{MMSE}(\xi,\rho)$ with respect to ρ is negative, after some tedious but straightforward calculations, can be shown to be equivalent to the following inequality

$$\rho\frac{2\xi^2+1}{\xi^2+1} \geq \tanh\left(\rho\frac{\xi^2(1+\xi^2)}{\xi^4+(1-\rho^2)(2\xi^2+1)}\right). \tag{23}$$

Using the fact that for any positive z we have $1 \geq \tanh(z)$, we can show that for $\rho \geq (\xi^2+1)/(2\xi^2+1)$ inequality (23) is true because

$$\rho\frac{2\xi^2+1}{\xi^2+1} \geq 1 \geq \tanh\left(\rho\frac{\xi^2(1+\xi^2)}{\xi^4+(1-\rho^2)(2\xi^2+1)}\right). \tag{24}$$

On the other hand for $\rho^2 \leq (\xi^2+1)/(2\xi^2+1)$, and since for $z \geq 0$ we have $z \geq \tanh(z)$, the inequality in (23) is again true because

$$\rho\frac{2\xi^2+1}{\xi^2+1} \geq \rho \geq \tanh(\rho) \geq \tanh\left(\rho\frac{\xi^2(1+\xi^2)}{\xi^4+(1-\rho^2)(2\xi^2+1)}\right). \tag{25}$$

And we complete the proof by noting that the two intervals of ρ, where (23) was shown to be true, completely cover the interval [0 1] due to the fact that $(\xi^2+1)/(2\xi^2+1) \leq 1$. ■

IV Conclusion

The results described in this paper have been pursued primarily out of theoretical interest. The significant practical advantages of the MMSE detector over the decorrelator and the matched filter would likely outweigh any performance disadvantage revealed here, inasmuch as the range of parameters for which the performance disadvantages arise are somewhat at the extremes for practical systems. Nevertheless, these results do provide some cautionary guidance concerning the relative merits of linear multiuser detectors.

It should be noted that further results can be obtained in this general area. For example, an interesting question arises as to the relative performance of the MF and MMSE detectors under the condition of perfect power control. One might conjecture that the MMSE detector would be superior in this case. However, counterexamples have been found that show this conjecture to be false in the cases of $K = 3$ and $K = 5$. These and additional related results will be included in future works on this subject.

References

[1] M. Honig and H. V. Poor, "Adaptive Interference Suppression," Chapter 2 in *Wireless Communications: Signal Processing Perspectives*, H. V. Poor and G. W. Wornell, Eds., (Prentice-Hall: Upper Saddle River, NJ, 1998), pp. 64 - 128.

[2] H.V. Poor and S. Verdú, "Probability of error in MMSE multiuser detection," *IEEE Trans. Inform. Theory*, vol. 43, no. 3, pp. 858-871, May 1997.

[3] S. Verdú, *Multiuser Detection*, (New York: Cambridge University Press, 1998)

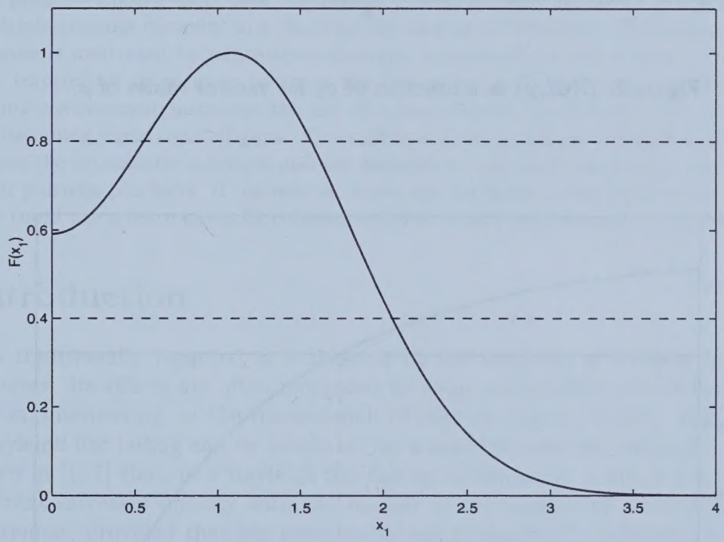


Figure 1: Typical form of the function $F(x_1)$

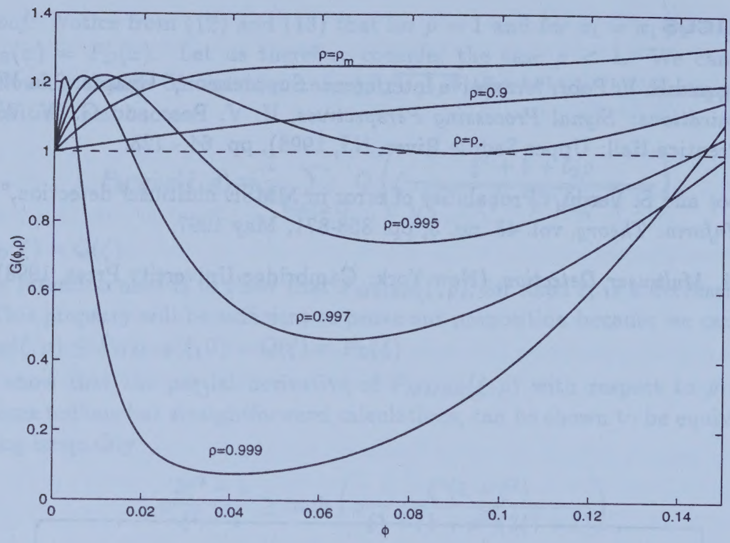


Figure 2: $G(\phi_2, \rho)$ as a function of ϕ_2 for various values of ρ.

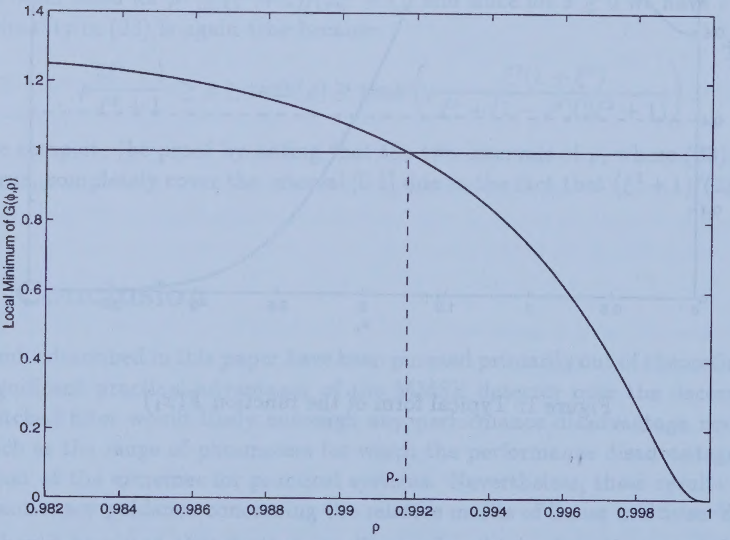


Figure 3: Local minimum of $G(\phi_2, \rho)$ as a function of ρ.

Space-Time Modulation Scheme for Unknown Rayleigh Fading Environments

BERTRAND M. HOCHWALD   THOMAS L. MARZETTA

Mathematical Sciences Research Center
Bell Laboratories, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974
{hochwald, tlm}@bell-labs.com

Abstract

We propose *unitary space-time modulation* for exploiting transmitter and receiver multiple-antenna diversity in a Rayleigh flat-fading environment. This space-time scheme is motivated by information-theoretic considerations and does not require the transmitter or receiver to know the fading environment. Not knowing the fading environment mandates the use of a noncoherent receiver and new criteria for designing signal constellations. Our matrix-valued signals are orthogonal in time across the transmitter antennas and are designed to minimize the singular values of their pairwise products. If the receiver knew the Rayleigh fading coefficients, then one could use a more powerful coherent receiver whose performance we compare.

1 Introduction

Fading is traditionally regarded as a nuisance by the designers of wireless communications systems. Its effects are often mitigated by some combination of differential phase modulation, interleaving, or the transmission of pilot or training signals. But, paradoxically, Rayleigh flat fading can be beneficial for a multiple-antenna communication link. It is shown in [1, 7] that, in a Rayleigh flat-fading environment, a link has a theoretical capacity that increases linearly with the smaller of the number of transmitter and receiver antennas, provided that the complex-valued propagation coefficients between all pairs of transmitter and receiver antennas are statistically independent and known to the receiver.

However, learning the fading coefficients becomes increasingly difficult as either the fading rate or number of transmitter antennas increases. In an effort to increase channel capacity or lower error probability, it is accepted practice to increase the number of transmitter antennas (thereby gaining "diversity" [5]). But increasing the number of transmitter antennas increases the required training interval and reduces the available time in which data may be transmitted before the fading coefficients change. At vehicle speeds of 60 miles/hour, a mobile operating at 1.9 GHz has a fading coherence interval of about 3 ms, which for a symbol rate of 30 kHz, corresponds to a fresh fade every 50–100 symbol periods. Assuming approximately ten training symbols per transmitter antenna, the coefficients for at most five antennas can be learned before a fresh fade occurs.

Motivated by these considerations, we used Shannon theory in [2] to analyze multiple-antenna links without imposing any training schemes and with no assumed knowledge of the random fading coefficients. The complex fading coefficients between all pairs of transmitter and receiver antennas were modelled as independent with uniformly distributed phases and Rayleigh distributed magnitudes. The fading coefficients were piecewise constant over fixed time intervals, with channel coding performed over many such independent fading intervals. We showed that the channel capacity could not be increased by making the number of transmitter antennas greater than the length of the fading interval, and found that the capacity-attaining signals had considerable structure. However, we did not explicitly address the problems of modulation and channel coding. In this paper, we use the structure derived in [2] to motivate a particular space-time modulation scheme.

The information-theoretic results in [2] suggest a signal constellation comprising complex-valued signals that are orthonormal with respect to time among the transmitter antennas. We call this signalling scheme *unitary space-time modulation.* When viewed as vector functions of time, the signals carry the message information entirely in their directions. We explain how to create, modulate, and demodulate unitary space-time signals on a multiple antenna link operating in Rayleigh flat fading. Throughout most of the paper, the propagation coefficients are assumed to be unknown to the receiver, but we also show how to use the modulation when the coefficients are known. When the receiver does not know the coefficients, no attempt to learn them is made. Some multiple-antenna coding issues for receivers that know the channel appear in [6]. We concentrate on modulation and constellation design, and do not address coding issues that lower error probability by adding redundancy. We focus, instead, on raw or uncoded signal and bit error probabilities. When combined with appropriate channel coding, our proposed signal constellations can theoretically attain a high fraction of the channel capacity.

2 Multiple-Antenna Link: Signal Model

Consider a communication link comprising M transmitter antennas and N receiver antennas that operates in a Rayleigh flat-fading environment. Each receiver antenna responds to each transmitter antenna through a statistically independent fading coefficient that is constant for T symbol periods. The received signals are corrupted by additive noise that is statistically independent among the N receivers and the T symbol periods. In complex baseband representation, during the T-symbol interval we transmit the signal $\{s_{tm},\ t = 1, \ldots, T,\ m = 1, \ldots, M\}$, and we receive the noisy signal $\{x_{tn},\ t = 1, \ldots, T,\ n = 1, \ldots, N\}$ related by the equation

$$x_{tn} = \sqrt{\rho/M} \sum_{m=1}^{M} h_{mn} s_{tm} + w_{tn}, \qquad t = 1, \ldots T, \ n = 1 \ldots N. \tag{1}$$

Here h_{mn} is the complex-valued fading coefficient between the mth transmitter antenna and the nth receiver antenna. The fading coefficients are constant for $t = 1, \ldots T$, and they are independent with respect to m and n and $\mathcal{CN}(0, 1)$ distributed (complex Gaussian). The transmitted signal has an average (over the M antennas) expected power equal to one,

$$\frac{1}{M}\sum_{m=1}^{M} \mathrm{E}\,|s_{tm}|^2 = 1, \qquad t = 1, \ldots, T. \tag{2}$$

The additive noise at time t and receiver antenna n is denoted w_{tn}, and is independent (with respect to both t and n), identically distributed $\mathcal{CN}(0,1)$. The quantities in the signal model (1) are normalized so that ρ represents the expected signal-to-noise ratio (SNR) at each receiver antenna, independently of the number of transmitter antennas. We assume that the realizations of h_{mn}, $m = 1, \ldots, M$, $n = 1, \ldots, N$ are not known to the receiver or transmitter.

We use matrix notation for the transmitted signal S ($T \times M$), and the received signal X ($T \times N$). Conditioned on S, the received signal X has independent and identically distributed columns (across the N antennas); at a particular antenna, the T received symbols are zero-mean circularly-symmetric complex Gaussian, with $T \times T$ covariance matrix $I_T + (\rho/M)SS^\dagger$, where I_T is the $T \times T$ identity matrix and $\dagger$ denotes conjugate-transpose. The received signal has conditional probability density,

$$p(X \mid S) = \frac{\exp\left(-\mathrm{tr}\left\{[I_T + (\rho/M)SS^\dagger]^{-1}XX^\dagger\right\}\right)}{\pi^{TN}\det^N[I_T + (\rho/M)SS^\dagger]}, \tag{3}$$

where "tr" denotes "trace."

We assume that the fading coefficients change to new independent realizations every T symbol periods. This piecewise constant fading process mimics, in a tractable manner, the behavior of a continuously fading process. Furthermore, it is a very accurate representation of many TDMA, frequency hopping, or block-interleaved systems [4]. Each channel use (consisting of a block of T transmitted symbols) is independent of every other. Thus, data can be transmitted reliably at any rate less than the channel capacity, where the capacity is the least upper bound on the mutual information between X and S, or

$$C = \sup_{p(S)} I(X;S),$$

subject to the average power constraint (2), and where

$$\begin{aligned} I(X;S) &= \mathrm{E}\log\frac{p(X \mid S)}{p(X)} \\ &= \int dS\, p(S) \int dX\, p(X \mid S) \log\left\{\frac{p(X \mid S)}{\int d\acute{S}\, p(\acute{S})p(X \mid \acute{S})}\right\}. \end{aligned} \tag{4}$$

3 Summary of Some Known Results

The conditional density (3) has considerable symmetry arising from the statistical equivalence of each time-sample and of each transmitter antenna. The special properties of the conditional density, in combination with the concavity of the mutual information functional, lead to some general conclusions [2] that are summarized here.

Theorem 1 *(Limit on number of transmitter antennas) For any coherence interval T and any fixed number of receiver antennas, the capacity obtained with $M > T$ transmitter antennas equals the capacity obtained with $M = T$ transmitter antennas.*

In what follows we assume that $M \leq T$.

Theorem 2 *(Structure of signal that achieves capacity) For large T, the signal matrix that achieves capacity can be written as $S = \sqrt{T}\Phi$, where Φ is an isotropically distributed $T \times M$ matrix whose columns are orthonormal.*

An *isotropically distributed* unit vector has a probability density that is unchanged when the vector is left-multiplied by any deterministic unitary matrix. Similarly, the isotropically distributed $T \times M$ matrix Φ obeys $\Phi^\dagger \Phi = I$, and has a density that is unchanged when it is left-multiplied by any $T \times T$ unitary matrix. In a natural way, Φ is the matrix counterpart of a complex scalar having unit magnitude and uniformly distributed phase; see [2] for more details.

4 Unitary Space-Time Modulation Attains Capacity for High SNR

Unitary space-time modulation defined

The previous section says that there is no point in making the number of transmitter antennas greater than the duration of the coherence interval. Furthermore, when the duration of the coherence interval is significantly greater than the number of transmitter antennas ($T \gg M$), setting $S = \sqrt{T}\Phi$ is virtually optimal. Taking our cue from these considerations, we define *unitary space-time modulation* to be the transmission of $S = \sqrt{T}\Phi$, where $\Phi^\dagger \Phi = I$. We now argue heuristically that unitary space-time modulation is optimal also for any fixed $T > M$, as $\rho \to \infty$. More formal arguments may be found in [3].

Suppose that there is $M = 1$ transmitter antenna and $N = 1$ receiver antenna, and let $s = v\Phi$, where v obeys $\mathrm{E}\, v^2 = T$ but is otherwise arbitrary, and consider the high-SNR received signal,

$$x \approx \sqrt{\rho} v h \Phi,$$

where x and Φ are T-dimensional vectors. The unit vector Φ, apart from its overall phase, can be determined very accurately from x, regardless of h. However, the scalar amplitude v cannot be determined so easily because it is multiplied by the unknown scalar fading coefficient h. Hence, when the SNR is high, transmitting information on Φ appears to be more profitable than transmitting on v. This suggests that we should simply set $v = \sqrt{T}$. Note that this argument applies only if $T > 1$. A similar intuitive argument also holds for multiple transmitters and receivers. When $S = \sqrt{T}\Phi$, all of the message information is carried on the directions of the orthonormal columns of Φ.

5 ML Receiver for Unitary Space-Time Modulation

We now consider maximum likelihood (ML) reception of a constellation of L signals employing unitary space-time modulation,

$$S_\ell = \sqrt{T}\Phi_\ell, \qquad \ell = 1, \ldots, L,$$

where $\{\Phi_\ell,\ \ell = 1, \ldots, L\}$ are $T \times M$ complex matrices satisfying $\Phi_\ell^\dagger \Phi_\ell = I$. We derive the ML receiver and its performance when H is unknown and, for comparison, when H

is known to the receiver (H is never known to the transmitter). It is customary to call the former receiver noncoherent and the latter receiver coherent.

5.1 Channel unknown to receiver

Maximum likelihood decoding becomes

$$\begin{aligned}
\Phi_{\mathrm{ml}} &= \arg \max_{\Phi_\ell \in \{\Phi_1,\ldots,\Phi_L\}} p(X \mid \Phi_\ell) \\
&= \arg \max_{\Phi_\ell \in \{\Phi_1,\ldots,\Phi_L\}} \frac{\exp\left(-\mathrm{tr}\left\{\left[I_T + (\rho T/M)\,\Phi_\ell \Phi_\ell^\dagger\right]^{-1} X X^\dagger\right\}\right)}{\pi^{TN} \det^N \left[I_T + (\rho T/M)\,\Phi_\ell \Phi_\ell^\dagger\right]} \\
&= \arg \max_{\Phi_\ell \in \{\Phi_1,\ldots,\Phi_L\}} \frac{\exp\left(-\mathrm{tr}\left\{\left[I_T - \frac{1}{1+M/\rho T}\Phi_\ell \Phi_\ell^\dagger\right] X X^\dagger\right\}\right)}{\pi^{TN}(1+\rho T/M)^{MN}} \\
&= \arg \max_{\Phi_\ell \in \{\Phi_1,\ldots,\Phi_L\}} \mathrm{tr}\,\{X^\dagger \Phi_\ell \Phi_\ell^\dagger X\}, \qquad (5)
\end{aligned}$$

where the matrix formulas $\det(I+AB) = \det(I+BA)$ and $(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$ are used. The ML receiver seeks to maximize the energy contained in the MN inner products that comprise $\Phi_\ell^\dagger X$.

Suppose now that $L = 2$, and Φ_1 and Φ_2 are transmitted with equal probability. The probability of decoding error is then

$$\begin{aligned}
P_e &= \frac{1}{2}\mathrm{P}(\mathrm{tr}\,\{X^\dagger \Phi_2 \Phi_2^\dagger X\} > \mathrm{tr}\,\{X^\dagger \Phi_1 \Phi_1^\dagger X\} \mid \Phi_1 \text{ transmitted}) \\
&\quad + \frac{1}{2}\mathrm{P}(\mathrm{tr}\,\{X^\dagger \Phi_1 \Phi_1^\dagger X\} > \mathrm{tr}\,\{X^\dagger \Phi_2 \Phi_2^\dagger X\} \mid \Phi_2 \text{ transmitted}). \qquad (6)
\end{aligned}$$

The probability of error given that Φ_1 is transmitted is equal to the probability of error given that Φ_2 is transmitted and is presented in the next theorem, whose proof may be found in [3].

Theorem 3 *(Two-signal error probability: H unknown) Suppose that two unitary space-time modulation signals Φ_1 and Φ_2 are transmitted with equal probability, and decoded with an ML receiver. Then the probability of error is*

$$\begin{aligned}
P_e &= \frac{1}{4\pi}\int_{-\infty}^{\infty} d\omega \frac{1}{\omega^2+1/4} \prod_{m=1}^{M} \left[\frac{1}{1+\frac{(\rho T/M)^2(1-d_m^2)(\omega^2+1/4)}{1+\rho T/M}}\right]^N \qquad (7) \\
&\leq \frac{1}{4\pi}\int_{-\infty}^{\infty} d\omega \frac{1}{\omega^2+1/4} \prod_{m=1}^{M} \left[\frac{1}{1+\frac{(\rho T/M)^2(1-d_m^2)}{4(1+\rho T/M)}}\right]^N \\
&= \frac{1}{2}\prod_{m=1}^{M} \left[\frac{1}{1+\frac{(\rho T/M)^2(1-d_m^2)}{4(1+\rho T/M)}}\right]^N, \qquad (8)
\end{aligned}$$

where $1 \geq d_1 \geq \ldots \geq d_M \geq 0$ are the singular values of the $M \times M$ matrix $\Phi_2^\dagger \Phi_1$, and (8) is the Chernoff bound. Decreasing any d_m decreases the total error probability.

For a single transmitter antenna ($M = 1$), d_1 is the magnitude of the inner product

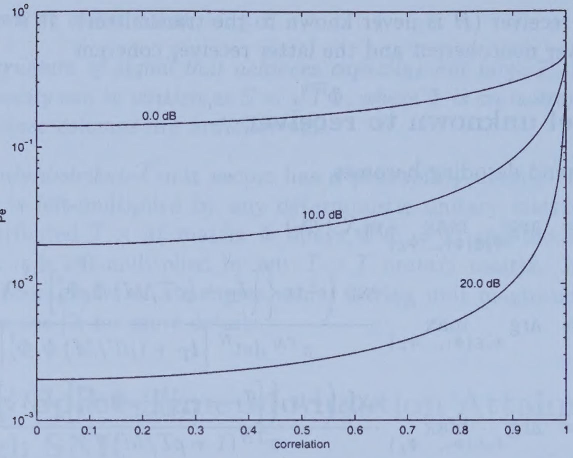


Figure 1: Two-signal probability of error vs. correlation d_1 for one transmitter and receiver antenna ($M = N = 1$), $T = 5$, and SNR=0, 10, 20 dB.

between Φ_1 and Φ_2. For multiple transmitter antennas, $d_1, \ldots, d_M$ represent the similarity between the subspaces spanned by the columns of Φ_1 and Φ_2. For given $d_1, \ldots, d_M$, the dependence of the probability of error on ρ and T is only through the product ρT.

Figure 1 shows the probability of error as a function of d_1 for $M = 1$ transmitter and $N = 1$ receiver antenna and $T = 5$ for SNR=0, 10, and 20 dB. We can see that reducing d_1 below approximately 0.4 does not reduce the error by much. Figure 2 superimposes the probability of error curves as a function of SNR for $M = 1$ and $M = 2$ transmitter antennas. For $M = 1$ the curves $d_1 = d$ are shown, and for $M = 2$ the curves $d_1 = d_2 = d$ are shown, where $d \in \{0.0, 0.8\}$. For SNR's greater than 5 dB, two transmitter antennas can have significantly lower error probability than one with the same total transmitted power. On the other hand, below approximately -2 dB, employing a second antenna with unitary space-time modulation actually increases the probability of error. This is not inconsistent with our claims that unitary space-time modulation is optimal for high SNR or large T. We conclude that when employing unitary space time modulation for given values of ρ, T, and N, there is an optimal number of transmitter antennas M that may be considerably smaller than T.

5.2 Channel known to receiver

We have justified unitary space-time modulation $S = \sqrt{T}\Phi$ on information-theoretic grounds for receivers that do not know the channel, when either $T \gg M$ or ρ is large. Surprisingly, we can also justify this modulation when $T \gg M$ and when the receiver knows the channel. When the receiver knows the channel, capacity is achieved by an S matrix composed of independent $\mathcal{CN}(0,1)$ random variables. In [2] it is argued that $S = \sqrt{T}\Phi$ (with Φ isotropically distributed) approaches, in distribution, a matrix of independent $\mathcal{CN}(0,1)$ random variables as $T \to \infty$. Hence, for T sufficiently large, unitary

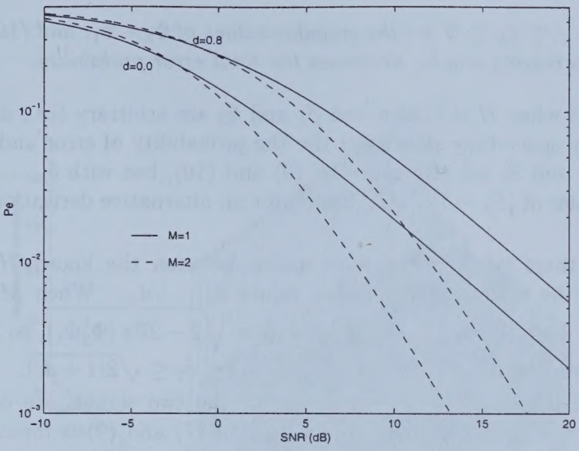


Figure 2: Two-signal probability of error vs. SNR for one ($M = 1$, solid curves), and two ($M = 2$, dashed curves) transmitter antennas, one receiver antenna ($N = 1$), $T = 5$, and $d = 0.0$, 0.8.

space-time modulation is nearly optimal, even when the channel is known. Knowledge of H, however, mandates different criteria for designing a signal constellation.

When H is known to the receiver (although still random),

$$p(X \mid S, H) = \frac{1}{\pi^{TN}} \exp\left(-\mathrm{tr}\left\{(X - \sqrt{\rho/M}SH)(X - \sqrt{\rho/M}SH)^\dagger\right\}\right),$$

and maximum likelihood decoding is

$$\Phi_{\mathrm{ml}} = \arg \min_{\Phi_\ell \in \{\Phi_1, \ldots, \Phi_L\}} \mathrm{tr}\{(X - \sqrt{\rho T/M}\Phi_\ell H)(X - \sqrt{\rho T/M}\Phi_\ell H)^\dagger\}.$$

The next theorem, which is also proven in [3], shows that the two-signal probability of error depends on the singular values of the $T \times M$ difference $\Phi_2 - \Phi_1$.

Theorem 4 *(Two-signal error probability: H known) Suppose that two unitary space-time modulation signals Φ_1 and Φ_2 are transmitted with equal probability, and decoded with an ML receiver that knows H perfectly. Then the probability of error, averaged over H, is*

$$\begin{aligned} P_e &= \frac{1}{4\pi}\int_{-\infty}^{\infty} d\omega \frac{1}{\omega^2 + 1/4} \prod_{m=1}^{M} \left[\frac{1}{1 + (\rho T/M)\delta_m^2(\omega^2 + 1/4)}\right]^N && (9) \\ &\le \frac{1}{4\pi}\int_{-\infty}^{\infty} d\omega \frac{1}{\omega^2 + 1/4} \prod_{m=1}^{M} \left[\frac{1}{1 + \frac{\rho T}{4M}\delta_m^2}\right]^N \\ &= \frac{1}{2}\prod_{m=1}^{M}\left[\frac{1}{1 + \frac{\rho T}{4M}\delta_m^2}\right]^N, && (10) \end{aligned}$$

where $2 \geq \delta_1 \geq \ldots \geq \delta_M \geq 0$ are the singular values of $\Phi_2 - \Phi_1$, and (10) is the Chernoff upper bound. Increasing any δ_m decreases the total error probability.

We note that when H is known and S_1 and S_2 are arbitrary (i.e., do not necessarily have the unitary space-time structure) the the probability of error and Chernoff bound for arbitrary S_1 and S_2 are still given by (9) and (10), but with $\delta_1, \ldots, \delta_M$ replaced by the singular values of $(S_2 - S_1)/\sqrt{T}$. See [6] for an alternative derivation of the Chernoff bound.

In general, there is no direct relationship between the known-H singular values $\delta_1, \ldots, \delta_M$, and the unknown-$H$ singular values $d_1, \ldots, d_M$. When $M = 1$, for example, we have $d_1 = |\Phi_2^\dagger \Phi_1|$ and $\delta_1 = ||\Phi_2 - \Phi_1|| = \sqrt{2 - 2\mathrm{Re}\,(\Phi_2^\dagger \Phi_1)}$, so for a given value of d_1, δ_1 can have the range of values $\sqrt{2(1-d_1)} \leq \delta_1 \leq \sqrt{2(1+d_1)}$.

For the special case $d_1 = \ldots = d_M = 0$ (the two signals are orthogonal), then $\delta_1 = \ldots = \delta_M = \sqrt{2}$, and a direct comparison of (7) and (9) is meaningful. For high SNR, the Chernoff bounds for H unknown (8) and H known (10) are then

$$P_e \lesssim \frac{1}{2}\left(\frac{4}{\rho T}\right)^{MN} \quad \text{(unknown)}, \qquad P_e \lesssim \frac{1}{2}\left(\frac{2}{\rho T}\right)^{MN} \quad \text{(known)},$$

which suggests that the probability of error is a factor of approximately 2^{MN} lower when the receiver knows H than when it does not.

We have seen that when H is known to the receiver, unitary space time modulation is a viable option for $T \gg M$. However, the maximum likelihood receivers for known H versus unknown H are considerably different, and so are the dependencies of probability of error on the signals. In the former we seek to maximize the singular values of $\Phi_2 - \Phi_1$, whereas in the latter we seek to minimize the singular values of $\Phi_2^\dagger \Phi_1$; these criteria are not compatible. Moreover, signal constellations for known H generally have to be larger than those for unknown H, reflecting the significantly higher channel capacity and lower error probability. When H is known, signals are distinguishable that would otherwise be indistinguishable if H were unknown, including antipodal pairs $\pm S$, as well as signals whose columns are permuted with respect to one another.

To demonstrate the viability of unitary space-time modulation, we generated a constellation of $L = 32$ signals for $T = 5$ (one bit per channel use), whose maximum correlation d_1 is 0.515. Figure 3 illustrates the correlations between the members of the constellation, $\Phi_1, \ldots, \Phi_{32}$. A constellation of $L = 32$ signals for $M = 2$ transmitter antennas was also generated, and Figure 4 compares the simulated bit error performance for a receiver that does not know the propagation coefficients. We see that the bit error probability decreases approximately as $1/\rho^2$ for high SNR with two antennas, versus approximately as $1/\rho$ with one antenna. No attempt was made to assign the data bits to the unitary space-time signals optimally.

6 Conclusions

Multiple element antenna arrays operating in Rayleigh flat fading can potentially sustain enormous data rates with moderate power in a narrow bandwidth. Our approach to this problem began with the premise that nobody knows the propagation coefficients and that the available transmission time should be spent sending message signals rather than

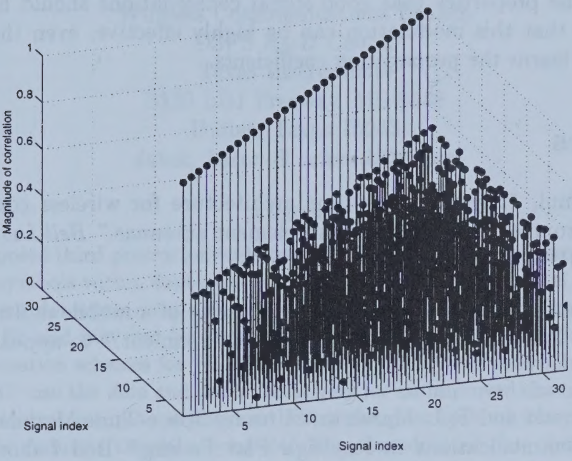


Figure 3: Magnitudes of correlations between $\Phi_1, \ldots, \Phi_{32}$ for $T = 5$. The diagonal entries with value 1.0 represent each signal correlated with itself.

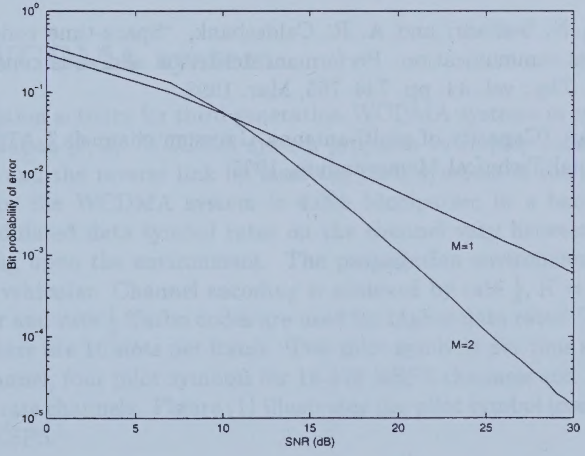


Figure 4: Performance of unitary space-time constellations for $M = 1$ versus $M = 2$ transmitter antennas for $T = 5$ as a function of SNR ρ, with one bit/channel use and channel unknown to receiver.

training signals. Information-theoretic considerations then led us to unitary space-time modulation. We have derived performance criteria for unitary space-time modulation and indicated the properties that good signal constellations should have. Preliminary results indicate that this modulation can be highly effective, even though the receiver never explicitly learns the propagation coefficients.

References

[1] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs. Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[2] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in a Rayleigh flat-fading environment," to appear in *IEEE Trans. Info. Thy.*

[3] B. M. Hochwald and T. L. Marzetta, "Unitary Space-Time Modulation for Multiple-Antenna Communications in Rayleigh Flat Fading," Bell Laboratories Technical Memorandum, 1998; also submitted to *IEEE Trans. Info. Thy.*

[4] L. H. Ozarow, S. Shamai (Shitz), and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Tech.*, vol. 43, pp. 359–378, May 1994.

[5] N. Seshadri and J. H. Winters, "Two signalling schemes for improving the error performance of frequency-division-duplex (FDD) transmission systems using transmitter antenna diversity," *Int. J. Wire. Inf. Net.*, vol. 1, no. 1, pp. 49–59, 1994.

[6] V. Tarokh, N. Seshadri and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inf. Thy.*, vol. 44, pp. 744–765, Mar. 1998.

[7] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," AT&T Bell Laboratories internal Technical Memorandum, 1995.

Iterative Channel Estimation for Wideband CDMA Systems

Anand G. Dabak, Srinath Hosur & Tim Schmidl
Wireless Communications Branch
DSPS R&D Center
Texas Instruments
8330 LBJ Freeway, MS 8368
Dallas, Texas 75243
dabak, hosur & schmidl@ti.com

Abstract

The proposed third generation wide band CDMA (WCDMA) systems periodically insert pilot symbols within data symbols to aid the receiver in channel estimation for doing coherent detection on the downlink. The expected Doppler frequencies range from 5-400 Hz and may not be known exactly to the receiver. Previously reported channel estimation schemes for WCDMA employ pilot symbol interpolation. We propose to jointly use the *data* and *pilots* iteratively for an improved channel estimation. An initial channel estimate is obtained using the pilot symbol interpolation which is used to demodulate the data symbols. The demodulated data symbols are then fed back and removed from the received input signal. A new channel estimate is obtained by a simple interpolation over the resulting signal. Two iterations of the above procedure are required for convergence of the improved channel estimate. We show that even with an error feedback of 14%, the iterative channel estimator (ICE) gives an improved $\frac{E_b}{N_0}$ gain of between 0.3-1.3 dB depending upon the Doppler rate. The increased complexity of the iterative channel estimator is about 3.2, 6.4 MIPS for the 32, 64 KSPS channels, a nominal increase for current DSP's. The ICE is robust in the sense, exact knowledge of the Doppler frequency is not required for the receiver.

1 The WCDMA system

The standardization activity for third generation WCDMA systems in currently under way in Europe and Japan [2, 3]. WCDMA system proposes to employ coherent demodulation for the forward and the reverse link by inserting pilot symbols in the data symbols. The chipping rate for the WCDMA system is 4.096 Mchips/sec in a bandwidth of 5 MHz. The QPSK modulated data symbol rates on the channel vary between 8 KSPS to 4.096 MSPS depending upon the environment. The propagation environment could be indoor, pedestrian and vehicular. Channel encoding is achieved by rate $\frac{1}{3}$, K = 9 for data rates 32 KSPS and lower and rate $\frac{1}{3}$ Turbo codes are used for higher data rates. The frame period is 10 msec. and there are 16 slots per frame. Two pilot symbols per time slot are inserted for the 8 KSPS channel, four pilot symbols for 16-128 KSPS channels and eight pilot symbols for higher data rate channels. Figure (1) illustrates the pilot symbol insertion for data rates 16, 32 and 64 KSPS.

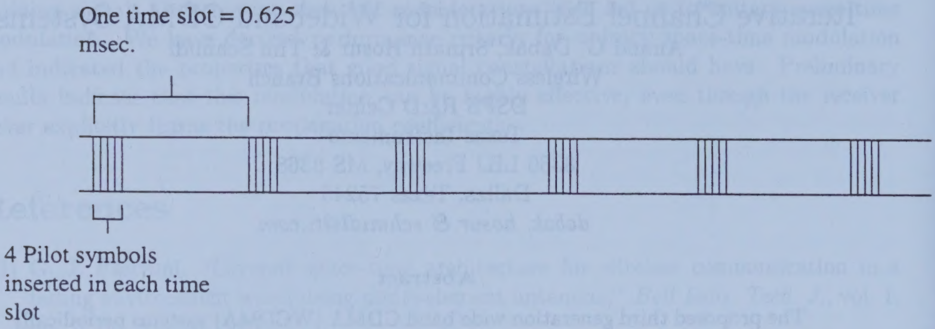

One time slot = 0.625 msec.
4 Pilot symbols inserted in each time slot

Figure 1: The 4 pilot symbol insertion for the WCDMA standard is shown for 16, 32 and 64 KSPS channels.

1.1 Iterative Channel Estimation for the WCDMA downlink

Previously reported channel estimation for WCDMA [1] proposes to employ pilot symbol interpolation (figure (2)). To do the channel estimation for the shaded portion in figure (2), the mobile uses a total of $2N$ sets of pilot symbols, N of them being the previous sets of pilot symbols and N from the future. The channel model is assumed to be Rayleigh fading. Had the mobile known the exact Doppler frequency, the Wiener filter [5] would be the optimal minimum mean squared estimate (MMSE) for the channel. However, *had* the receiver *known* the modulation of the data symbols, it could have improved upon the channel estimates by using the channel values observed during the data symbol period. This is possible by feeding back the data symbols at the output of the demodulator to the received signal at the input and removing the data modulation from the received signal. We can repeat this procedure multiple times to further improve upon the channel estimate. Figures (3) and (4) illustrate this idea of iterative channel estimation (ICE).

The idea of ICE is similar to the idea of other *joint pilot and data* aided channel estimation as given in [4, 8, 7]. However, to our knowledge no analysis has been persented previously for the WCDMA case.

2 Performance of ICE

We now try to quantify in terms of loss in signal to noise ratio or $\frac{E_b}{N_0}$ the performance of the ICE and compare it to the performance of channel estimation using only the pilots. Let the received signal after despreading be denoted by;

$$r_i^j = a_i^j d_i + n_i^j;\ n_i^j \sim N(0,\sigma_n)\ j = 1,\cdot,\cdot,L \qquad (1)$$

In the above equation, r_i^j denotes the despread signal of the j^{th} multi-path of the i^{th} symbol. Similarly a_i^j, n_i^j denotes the channel Rayleigh fading coefficient and the additive Gaussian

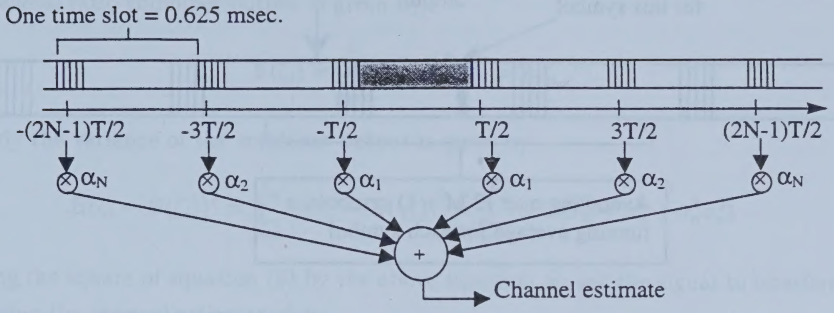

One time slot = 0.625 msec.
-(2N-1)T/2
-3T/2
-T/2
T/2
3T/2
(2N-1)T/2
α_N
α_2
α_1
α_1
α_2
α_N
+
Channel estimate

Figure 2: Channel estimation by pilot symbol interpolation as reported in [1].

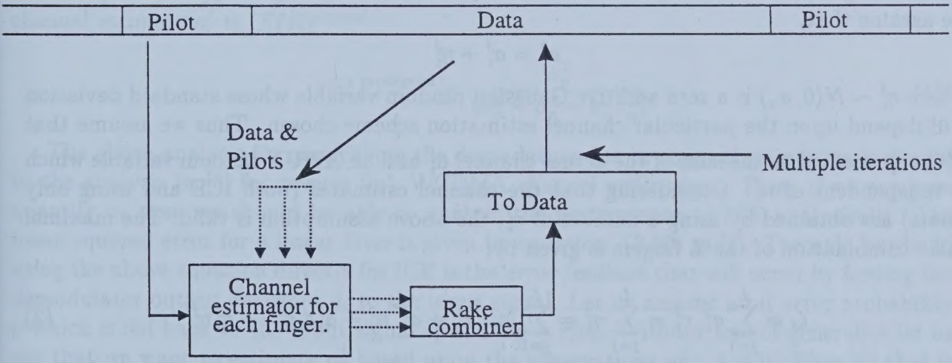

Pilot
Data
Pilot
Data &
Pilots
Multiple iterations
To Data
Channel
estimator for
each finger.
Rake
combiner

Figure 3: A block diagram of the proposed ICE scheme is shown. In the first iteration, the channel estimation is performed using the interpolation technique shown in figure(2). The demodulator output using this channel estimator is then used to remove the modulation of the data symbols from the incoming signal. A new channel estimate is performed using this signal by averaging over $(2M + 1)$ symbols (figure(4)). This procedure is repeated a second time.

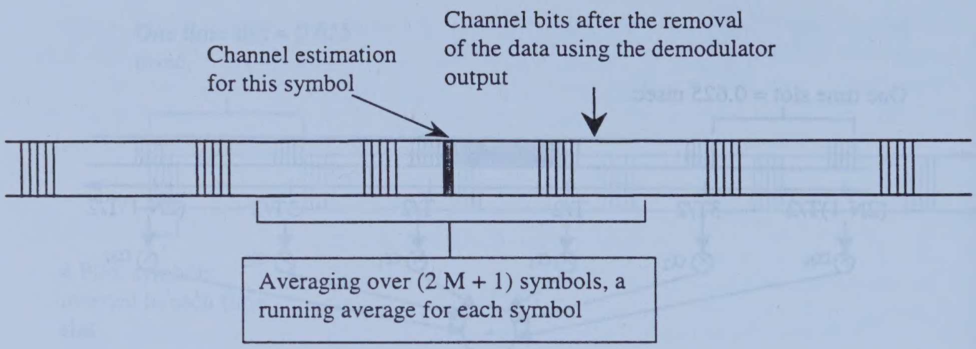

Channel bits after the removal of the data using the demodulator output
Channel estimation for this symbol
Averaging over (2 M + 1) symbols, a running average for each symbol

Figure 4: Channel estimation employed for iterations > 1 is shown. A long filter of length $(2M+1)$ using both the pilot and data symbols is used for channel estimation.

---

noise of the j^{th} multi-path of the i^{th} symbol. The total number of multi-paths is L. For the WCDMA system, L is typically equal to 4 in vehicular environment. Similarly let $\hat{a_i^j}$ denote the channel estimate (obtained either by ICE or using only the pilots) of a_i^j. Then we assume that,

$$\hat{a_i^j} = a_i^j + \eta_i^j$$

where $\eta_i^j \sim N(0, \sigma_\eta)$ is a zero additive Gaussian random variable whose standard deviation will depend upon the particular channel estimation scheme chosen. Thus we assume that $\hat{a_i^j}$ is represented as the sum of the actual channel a_i^j and an AWGN random variable which is independent of n_i^j. Considering that the channel estimates (both ICE and using only pilots) are obtained by using a number of r_i^j, the above assumption is valid. The maximal ratio combination of the L fingers is given by;

$$\zeta_i = \sum_{j=1}^{L} \hat{a_i^j}^* r_i^j = \sum_{j=1}^{L} \gamma_i^j = \sum_{j=1}^{L} |a_i^j|^2 d_i + a_i^{j*} n_i^j + a_i^j d_i \eta_i^{j*} + n_i^j \eta_i^{j*} \tag{2}$$

where we have denoted $\gamma_i^j = \hat{a_i^j}^* r_i^j$. We now get

$$E(\gamma_i^j / d_i, a_i^j) = |a_i^j|^2 d_i \tag{3}$$

and

$$\begin{aligned} E(\gamma_i^j - E(\gamma_i^j / d_i, a_i^j))^2 &= E\left(a_i^{j*} n_i^j + a_i^j d_i \eta_i^{j*} + n_i^j \eta_i^{j*}\right)^2 && (4) \\ &= |a_i^j|^2 \sigma_n^2 + |a_i^j|^2 \sigma_\eta^2 + \sigma_n^2 \sigma_\eta^2 && (5) \end{aligned}$$

Substituing the above statistics back into equation (2) and now averaging over all a_i^j and assuming that the signal power $E\left(|a_i^j|^2\right) = 1$ for each path, we get that the statistics of the maximal ratio combiner output is given by;

$$E(\zeta_i) = \sum_{j=1}^{L} E\left(|a_i^j|^2\right) d_i = L d_i \tag{6}$$

Similarly the variance of the combiner output is given by;

$$E(\zeta_i - E(\zeta_i))^2 = \sum_{j=1}^{L} E(\gamma_i^j - E(\gamma_i^j))^2 = \sum_{j=1}^{L} = L(\sigma_n^2 + \sigma_\eta^2 + \sigma_n^2\sigma_\eta^2) \tag{7}$$

Dividing the square of equation (6) by the above equation we get the signal to interference ratio using the channel estimate $\hat{a}_i^j$ is

$$SIR = \frac{L}{(\sigma_n^2 + \sigma_\eta^2 + \sigma_n^2\sigma_\eta^2)} \tag{8}$$

On the other hand, the signal to interference ratio of an ideal maximal ratio combiner would be $(SIR)_{ideal}$

$$(SIR)_{ideal} = \frac{L}{\sigma_n^2} \tag{9}$$

Diving equations (8) by (9) we the degradation in signal to interference ratio using non-ideal channel estimate $\hat{a}_i^j$ is $(SIR)^{degrad}$

$$(SIR)^{degrad} = \frac{\sigma_n^2}{(\sigma_n^2 + \sigma_\eta^2 + \sigma_n^2\sigma_\eta^2)} \tag{10}$$

The above analysis for quantifying the degradation due to channel estimation is similar to the analysis in [6] for reverse link WCDMA channel estimation. Thus, if we can now quantify σ_η then we should be able to analyze the performance of ICE analytically. The mean squared error for a linear filter is given by equation (3.20) in [5]. The only hurdle in using the above equation directly for ICE is the *error feedback* that will occur by feeding the demodulator output decisions $\hat{d}_i$ to the input signal. Let us assume a bit error probability p which is fed back to the input signal r_i^j and $w_i^j = r_i^j \hat{d}_i$. Without loss of generality let us say that we want to estimate a_0^j based upon the observations $w_i^j;\ j \neq 0$. Then all that is needed to compute σ_η using equation (3.20) in [5] is

$$E(a_0^j w_i^j) = (1-p)E(a_0^j a_i^j) - pE(a_0^j a_i^j) = (1-2p)E(a_0^j a_i^j) \tag{11}$$

and for $k \neq i$

$$E(w_i^j w_k^j) = (p^2 + (1-p)^2)E(a_i^j a_k^j) - 2p(1-p)E(a_i^j a_k^j) = (1-2p)^2 E(a_i^j a_k^j) \tag{12}$$

and for $k = 1$

$$E(w_i^j w_i^j) = E(|a_i^j|)^2 + \sigma_n^2 \tag{13}$$

3 Simulation Results

Using the $\frac{E_b}{N_0}$ degradation in equation (10) the performance of ICE was compared to the performance of channel estimation using only the pilots. The details of the simulations are give in the following table;

Channel symbol rate (Ksps)	32	64
N	3, Doppler 5, 80 Hz	3, Doppler 5, 80
N	2, Doppler 200, 320 Hz	Doppler 200, 320
M	40, Doppler 5, 80 Hz,	80, Doppler 5, 80
M	20, Doppler 200, 320 Hz	40, Doppler 200, 320
Fingers L	4	4

Table 1

Using the above parameters, figure (5) compares the performance of ICE with that using only the pilots for channel estimation. The operating point for the WCDMA system is around an uncoded ber of about 0.14 which implies an uncoded $\frac{E_b}{N_0}$ of about -1.0 dB in figure (5). In order to confirm the analysis for the performance of ICE actual Monte Carlo simulations were performed and the performance of ICE was compared to that using only pilots for channel estimation. The operating point chosen was an uncoded ber of 0.14. The results of the Monte Carlo simulations are given in the following table;

Channel estimation	Pilots interpolation	ICE	Gain of ICE (dB)
Doppler = 5	-1.3	-1.6	0.3
Doppler = 80	-1.0	-1.3	0.3
Doppler = 200	-0.4	-0.9	0.5
Doppler = 320	0.4	-0.4	0.8

Table 2: The required $\frac{E_b}{N_0}$ obtained using Monte Carlo simulations, for achieving an uncoded target ber of 0.14 is shown for 32 KSPS channel symbol rate. We can see that using ICE gives a gain of about 0.3-0.8 dB.

Channel estimation	Pilots interpolation	ICE	Gain of ICE (dB)
Doppler = 5	-1.3	-1.8	0.5
Doppler = 80	-1.0	-1.5	0.5
Doppler = 200	-0.4	-1.2	0.8
Doppler = 320	0.5	-0.8	1.3

Table 3: The required $\frac{E_b}{N_0}$ obtained using Monte Carlo simulations, for achieving an uncoded target ber of 0.14 is shown for 64 KSPS channel symbol rate. We can see that using ICE gives a gain of about 0.5-1.3 dB.

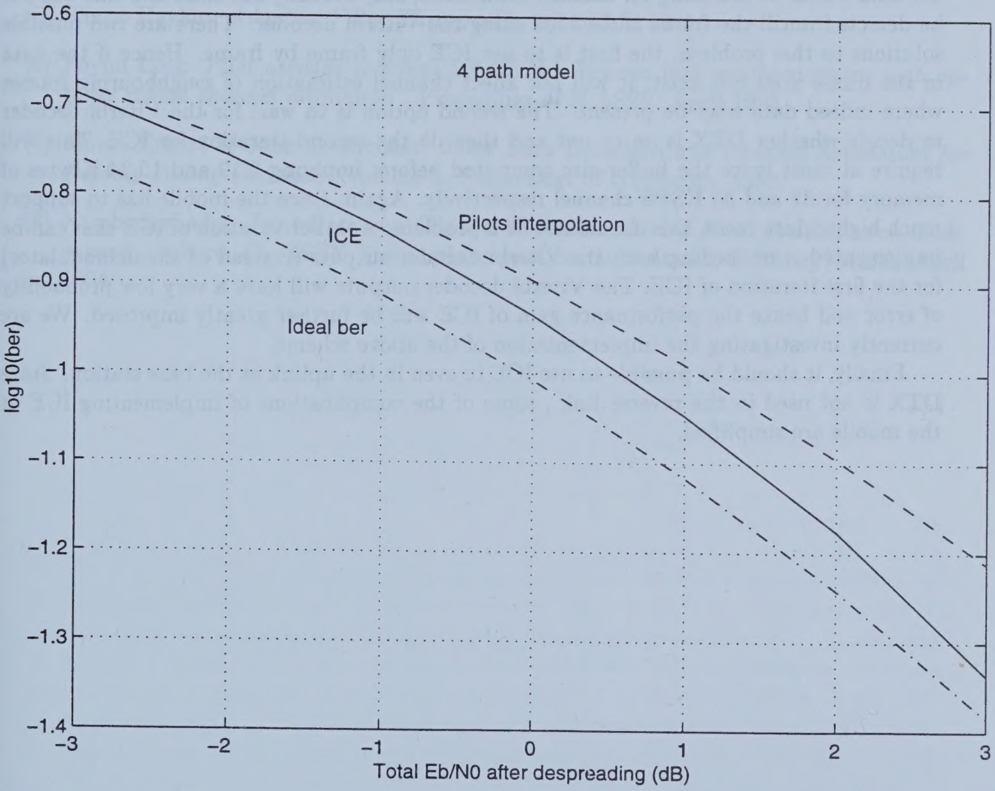


Figure 5: The bit error rate (ber) at the output of the maximal ratio combiner (MRC) is plotted for 2 iterations of ICE and channel estimation using only the pilots for the 64 KSPS channel symbol rate and Doppler of 200 Hz.

4 Implementation Issues

Each iteration of ICE needs $2L$ extra complex adds and an extra MRC for channel estimation and demodulation of each symbol. Assuming that each complex add is 2 real operations and each MRC is 6 real operations, we need $10 \times 4 \times 32000 \times 2 = 2.56$ MIPS extra for 32 KSPS and $10 \times 4 \times 64000 \times 2 = 5.12$ MIPS extra for the 64 KSPS channel. Also, the mobile has to buffer 1 full frame of data which corresponds to $32000 \times 10 \times 10^{-3} \times 2 \times 4 = 2.56$ Kbytes of data for the 32 KSPS channel and $64000 \times 10 \times 10^{-3} \times 2 \times 4 = 5.12$ Kbytes of data for the 64 KSPS channel. Using the pilot symbol interpolation for data rates 128 KSPS and higher will require the existence of a large buffer in the memory. Hence supporting ICE for lower data rates, 8, 16, 32 and 64 KSPS should not be a problem. Another implementation issue is that of discontinuous transmission (DTX) that is used in WCDMA systems. Thus the data which we are using for channel estimation, may actually not exist and this will not be detected untill the frame is decoded using the Viterbi decoder. There are two possible solutions to this problem, the first is to use ICE only frame by frame. Hence if the data in the frame does not exist, it will not affect channel estimation of neighbouring frames where indeed data may be present. The second option is to wait for the Viterbi decoder to decide whether DTX is on or not and then do the second iteration for ICE. This will require at most twice the buffer size computed before, implying 5.12 and 10.24 Kbytes of memory for 32 and 64 KSPS channel respectively. Again, since the mobile has to support much higher data rates, this should not be a problem. Another varation of ICE that can be implemented is by feeding back the Viterbi decoder outputs (instead of the demodulator) for the first iteration of ICE. The Viterbi decoder outputs will have a very low probability of error and hence the performance gain of ICE will be further greatly improved. We are currently investigating the implementation of the above scheme.

Finally, it should be possible to use ICE in even in the uplink at the base station. Since DTX is not used in the reverse link , some of the complications of implementing ICE at the mobile are simplified.

References

[1] H. Andoh K. Higuchi K. Okawa M. Sawahashi & F. Adachi. *Experiments on Pilot-Symbol Assisted Coherent Rake Receiver for W-CDMA Mobile Radio* . Technical Report of IEICE, April 1998.

[2] ARIB. *Specification of Air-Interface for 3G Mobile System, Version 0.5, April 1998* .

[3] ETSI. *Tdoc SMG2 359/97, December 1997* .

[4] Kuor-Hsin Chang & Costas N. Georghiades. *Iterative Joint Sequence and Channel Estimation for Fast Time-Varying Itersymbol Interference Channels* . Procedings of ICC, Seattle, 1995.

[5] Simon Haykin. *Adaptive Filter Theory* . Prentice Hall, 1986.

[6] Fuyun Ling. *Pilot assisted coherent DS-CDMA Reverse link communications with optimal robust channel estimation* . Procedings of ICASSP, April 1997.

[7] Murat Erkurt & John G. Proakis. *Joint Data Detection and Channel Estimation for Rapidly Fading Channels* , volume 2. Procedings of Gloebcom, 1992.

[8] Nambi Seshadri. *Joint Data and Channel Estimation Using Blind Trellis Search Techniques* , volume 42. IEEE Transactions on Communications, February/March/April 1994.

An Inexact Hybrid Projection-Proximal Augmented Lagrangian Algorithm

Carlos Humes Junior*
Dept. of Computer Science
University of São Paulo - Brazil
humes@ime.usp.br

Paulo José da Silva e Silva†
Dept. of Applied Mathematics
University of São Paulo - Brazil
rsilva@ime.usp.br

B. F. Svaiter‡
Instituto de Matemática Pura e Aplicada
Rio de Janeiro - Brazil
benar@impa.br

Abstract

In this work, we use a recent result from Solodov and Svaiter on hybrid projection-proximal point methods [13] to derive an algorithm to find a Kuhn-Tucker pair of a convex programming problem. The algorithm is a variation of the proximal augmented Lagrangian, where at each iteration we minimize *approximately* a function $\phi_k(\cdot)$ defined as the classical augmented Lagrangian added to a primal regularization. The main feature of the algorithm is that it does not require increasing accuracy in the solution of the unconstrained subproblems. We also show that the convergence is Q-linear under strong second order assumptions.

Preliminary computational experiments are also presented.

1 Introduction

Rockafellar presented several applications of the proximal point algorithm [9, 8, 11, 7] to convex optimization [10], in particular the *proximal method of multipliers*, presented in section 5. This method is the proximal point algorithm applied to find a zero of the maximal monotone operator associated to the Lagrangian saddle function. Each step of this algorithm requires an inexact minimization with increasing relative accuracy. It was proved that the sequence generated converges to a saddle point of the Lagrangian function and that the rate of convergence is Q-linear under strong second order assumptions.

In this work, based on a recent result from Solodov and Svaiter on hybrid projection-proximal methods [13], we derive a variation of Rockafellar's proximal method of multipliers that has the same convergence properties, but requires only constant relative accuracy. Both methods coincide if exact minimization is done at each step. Some computational results are presented.

---

*The research of this author is supported by PRONEX-Optimization

†The research of this author is supported by FAPESP, Grant 96/09939-0.

‡The research of this author is supported by CNPq, Grant 301200/93-9(RN) and by PRONEX-Optimization.

2 The Algorithm

2.1 Definitions and notation

Consider the problem

$$\begin{cases} \min & f(x) \\ s.t. & g(x) \leqq 0 \\ & x \in \mathbb{R}^n, \end{cases} \tag{P}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $g : \mathbb{R}^n \mapsto \mathbb{R}^m$. We assume throughout this paper that $f(\cdot)$ and each component of $g(\cdot)$ are convex and differentiable[1]. Moreover, we assume that (P) and its dual problem have solutions.

The extended Lagrangian function of (P) is:

$$l(x,y) := \begin{cases} f(x) + \sum_{j=1}^{m} y_j g_j(x), & \text{if } y \in \mathbb{R}_+^m \\ -\infty, & \text{otherwise.} \end{cases}$$

Saddle points of the Lagrangian are the Kuhn-Tucker pairs of (P) and therefore the pairs of solutions of (P) and its dual.

Associated to the Lagrangian, $l(\cdot,\cdot)$, we have the maximal monotone operator $T_l : (x,y) \mapsto \{(u,v) \mid (v,-u) \in \partial l(x,y)\}$, *i.e.*:

$$T_l(x,y) := \begin{bmatrix} \nabla f(x) + \sum_{j=1}^{m} y_j \nabla g_j(x) \\ -g(x) + \mathrm{N}_{\mathbb{R}_+^m}(y) \end{bmatrix},$$

where

$$\mathrm{N}_{\mathbb{R}_+^m}(y)_j = \begin{cases} 0, & \text{if } y_j > 0; \\ (-\infty, 0], & \text{if } y_j = 0; \\ \emptyset, & \text{if } y_j < 0. \end{cases}$$

In [10], to find a zero of $T_l(\cdot,\cdot)$, one generates a sequence $(x^k, y^k)_{k\in\mathrm{N}}$ where x^{k+1} is an approximate unconstrained minimum of the function

$$\phi_k(x) := f(x) + P(g(x), y^k, c^k) + (1/2c^k)\,\|x - x^k\|^2, \tag{1}$$

and

$$P(w,y,c) := (1/2c)\sum_{j=1}^{m}\left[(y_j + cw_j)_+{}^2 - (y_j)^2\right].^{\dagger} \tag{2}$$

Then one defines $y^{k+1} := \nabla_w P(g(x^{k+1}), y^k, c^k)$.[3]

---

[1] It should be clear to reader that we could drop the differentiability assumptions by assuming that the functions are real valued in convenient domains and using subdifferentiability. For the sake of exposition's simplicity, we prefer to consider only smooth problems.

[†] As usual, a_+ denotes $\max\{0, a\}$. We shall use a_- for the corresponding minimum operation and also allow these operations to be applied componentwise to vectors.

[3] Observe that $P(\cdot, y, c)$ is Lipschitz continuously differentiable in w with Lipschitz constant c, but it is not twice differentiable (with respect to w) at $w = (-1/c)y$.

An approximate solution to the minimization of $\phi_k(\cdot)$ is considered acceptable if

$$\| \nabla\phi_k(x^{k+1}) \| \leqq \epsilon^k / c^k, \quad \sum_{i=0}^{\infty} \epsilon^k < \infty; \tag{A}$$

or

$$\| \nabla\phi_k(x^{k+1}) \| \leqq (\delta^k / c^k) \, \| (x^{k+1}, y^{k+1}) - (x^k, y^k) \|, \quad \sum_{i=0}^{\infty} \delta^k < \infty. \tag{B}$$

This error tolerance comes directly from the errors bounds demanded by the proximal point algorithm. In particular, an exact solution, x^{k+1} of the minimization of $\phi_k(\cdot)$ and the corresponding $y^{k+1} := \nabla_w P(g(x^{k+1}), y^k, c^k)$ would be a solution of the exact proximal step.

Rockafellar used the criteria (A) to ensure convergence and (B) to prove Q-linear convergence rate.

2.2 The Hybrid algorithm

In [13], Solodov and Svaiter introduce a variation of the proximal point algorithm, called hybrid projection-proximal point algorithm, with less stringent error. To achieve this, a projection onto a special hyperplane generated from the approximate solution of the proximal step must be done at each iteration.

In our particular case the algorithm becomes:

Hybrid Projection-Proximal Point Algorithm

1. *Initialization: Let* $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m_+$ *and* $\sigma \in [0, 1)$.
2. *Iteration: Given* (x^k, y^k) *and* $c^k > 0$.
   (a) *Inner step: Find* $(\tilde{x}^k, \tilde{y}^k)$ *and a* $v^k \in T_l(\tilde{x}^k, \tilde{y}^k)$*, such that*

$$\| v^k + 1/c^k((\tilde{x}^k, \tilde{y}^k) - (x^k, y^k)) \| \leqq \sigma \max\{\| v^k \|, 1/c^k \, \| (\tilde{x}^k, \tilde{y}^k) - (x^k, y^k) \|\}, \tag{3}$$

   (b) *Projection: If* $v^k = 0$*, or* $(\tilde{x}^k, \tilde{y}^k) = (x^k, y^k)$*, Stop.*
   Otherwise,

$$(x^{k+1}, y^{k+1}) := (x^k, y^k) - \frac{\langle v^k, (x^k, y^k) - (\tilde{x}^k, \tilde{y}^k) \rangle}{\| v^k \|^2} v^k. \tag{4}$$

■

Clearly, the projection step projects the point (x^k, y^k) onto the hyperplane that has v^k as normal and intercepts $(\tilde{x}^k, \tilde{y}^k)$.

As stated in the last section, the exact minimization of $\phi_k(\cdot)$, gives a solution of the proximal step, *i.e.* a pair $(\tilde{x}^k, \tilde{y}^k)$ that has a $v^k \in T_l(\tilde{x}^k, \tilde{y}^k)$ such that:

$$v^k + 1/c^k((\tilde{x}^k, \tilde{y}^k) - (x^k, y^k)) = 0.$$

Then, it is natural to use an approximate solution of this minimization problem to perform the inner step above. The main difficulty here is how to compute a good element in $T_l(\cdot, \cdot)$ that permits testing the inner acceptance criteria, (3), and to project, (4). This will be handled by the next two simple results:

Lemma 1. *Let* $\widetilde{x} \in \mathbb{R}^n$ *and* $y \in \mathbb{R}^m_+$ *and* $c > 0$. *Define*

$$\widetilde{y} := \nabla_w P(g(\widetilde{x}), y, c) \quad and \quad \nu := \big(y + cg(\widetilde{x})\big)_-,$$

then

$$\nu \in N_{\mathbb{R}} m(\widetilde{y}),$$

and

$$v := \begin{bmatrix} \nabla_x l(\widetilde{x}, \widetilde{y}) \\ -g(\widetilde{x}) + \frac{1}{c}\nu \end{bmatrix} \in T_l(\widetilde{x}, \widetilde{y}).$$

Proof. The result follows trivially, observing that $\nabla_w P(g(\widetilde{x}), y, c) = \big(y + cg(\widetilde{x})\big)_+$. ☐

The reason of choosing ν among the elements of $N_{\mathbb{R}}\mathrm{m}(\widetilde{x})$ becomes clear as:

Proposition 1. *Let* $\widetilde{x}^k \in \mathbb{R}^n$ *and* $y^k \in \mathbb{R}^m_+$ *and* $c^k > 0$. *Defining* $\widetilde{y}^k$, ν^k *and* v^k *as above it follows that:*

$$r^k := v^k + 1/c^k\big((\widetilde{x}^k, \widetilde{y}^k) - (x, y)\big) = \begin{bmatrix} \nabla\phi_k(\widetilde{x}^k) \\ 0 \end{bmatrix} \tag{5}$$

Proof. Clearly, $\nabla\phi_k(\widetilde{x}^k) = \nabla_x l(\widetilde{x}, \widetilde{y}) + 1/c^k(\widetilde{x}^k - x^k)$.

On the other hand

$$\begin{aligned} -g(\widetilde{x}^k) &+ (1/c^k)\nu^k + 1/c^k(\widetilde{y}^k - y^k) = \\ &= -g(\widetilde{x}^k) + 1/c^k \min\{y^k + c^k g(\widetilde{x}^k), 0\} + 1/c^k\big(\max\{y^k + c^k g(\widetilde{x}^k), 0\} - y^k\}\big) \\ &= 1/c^k\big(\min\{y^k, -c^k g(\widetilde{x}^k)\} + \max\{c^k g(\widetilde{x}^k), -y^k\}\big) \\ &= 0. \end{aligned}$$

This completes the proof. ☐

Now it's easy to present an implementable form of the hybrid projection-proximal algorithm, that we call the *inexact hybrid projection-proximal augmented Lagrangian.*

Inexact Hybrid Projection-Proximal Augmented Lagrangian Method

1. *Initialization: Let* $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m_+$ *and* $\sigma \in [0, 1)$.
2. *Iteration: Given* (x^k, y^k) *and* $c^k > 0$, *define* $P(\cdot, y^k, c^k)$ *as in equation (2) and* $\phi_k(\cdot)$, *as in equation (1).*
    (a) *Inner optimization: Find an approximate solution of the unconstrained optimization of* $\phi_k(\cdot)$, $\widetilde{x}^k$, *such that, if we define*

$$\widetilde{y}^k := \nabla_w P(g(\widetilde{x}^k), y^k, c^k)$$
$$v^k := \begin{bmatrix} \nabla_x l(\widetilde{x}^k, \widetilde{y}^k) \\ -g(\widetilde{x}^k) + 1/c^k\nu \end{bmatrix},$$

then

$$\|\nabla\phi_k(\widetilde{x}^k)\| \leqq \sigma \max\{\|v^k\|, 1/c^k\, \|(\widetilde{x}^k, \widetilde{y}^k) - (x^k, y^k)\|\}.$$

(b) Projection: If $v^k = 0$, or $(\tilde{x}^k, \tilde{y}^k) = (x^k, y^k)$, Stop.
Otherwise,

$$(x^{k+1}, y^{k+1}) := (x^k, y^k) - \frac{\langle v^k, (x^k, y^k) - (\tilde{x}^k, ty^k)\rangle}{\| v^k \|^2} v^k.$$

■

The main convergence theorems are:

Theorem 1. *If the problem (P) and its dual have solutions and the sequence of penalization parameters, $(c^k)_{k \in \mathbb{N}}$, is bounded away from zero; then the sequence generated by the inexact hybrid projection-proximal augmented Lagrangian converge to a pair of solution of these problems (a Kuhn-Tucker point).*

Proof. This is a corollary of the convergence of the inexact projection-proximal point presented in theorem 2.1 in [13]. □

Theorem 2. *Under the assumptions of theorem 1, if $T_l^{-1}(\cdot,\cdot)$ is Lipschitz continuous at the origin then the the convergence rate is at least Q-linear.*

Proof. This is a corollary of the rate of convergence of the hybrid projection-proximal algorithm, theorem 2.2 in [13]. □

Note that Lipschitz continuity of $T_l^{-1}(\cdot,\cdot)$ can be guaranteed under strong second order condition, please see discussion in [10], pp. 102-103.

It is important to stress that, to the authors knowledge, this is the *first* convergence result of an optimization method similar to the Augmented Lagrangian algorithm that does not require increasing accuracy, *i.e.*, the σ accuracy is held constant during the whole process. All the convergence results so far, asked the accuracy to decrease to zero [1, 2, 10].

3 Computational Experiments

In this section, we present some preliminary computational results to validate the applicability of the algorithm presented above. We also do some comparisons between our algorithm and the classical augmented Lagrangian with errors, as presented in Bertsekas [1].

3.1 The Tests Problems

We used the following convex test problems:

- From Hock and Schttikowski collection [5, 12]: problems 21, 28, 35, 51, 76, 215, 218, 224, 262, 268, 284, 315, 384;
- From Lasdon [6]: problems 1 and 3;
- The following small and simple problem:

$$\left\{ \begin{array}{ll} \min & \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ s.t. & (x_1 - 2)^2 + (x_2 - 2)^2 - 1 \leqq 0 \\ & (x_2 - 3)^2 + (x_3 - 3)^2 - 3 \leqq 0. \end{array} \right.$$

These are all small-scale problems, with up to 15 variables and 10 constraints. In general they are quadratic problems.

3.2 The algorithms implemented

We have implemented the hybrid projection-proximal augmented Lagrangian described above and the augmented Lagrangian with inexact minimization as described in [1], chapter 5. For the sake of completeness we present this algorithm:

Augmented Lagrangian with Inexact Minimization

1. *Initialization: Let* $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m_+$ *and* $(\epsilon^k)_{k\in\mathrm{N}}$ *any sequence converging to zero.*

2. *Iteration: Given* (x^k, y^k) *and* $c^k > 0$*, define* $P(\cdot, y^k, c^k)$ *as in equation (2), and define*

$$\varphi_k(x) := f(x) + P(g(x), y+, c^k).$$

   (a) *Inner optimization: Find an approximate solution of the unconstrained optimization of* $\varphi_k(\cdot)$*,* x^{k+1}*, such that*

$$\| \nabla \varphi_k(x^{k+1}) \| \leqq \epsilon^k / c^k \, \| \nabla_w P(g(x^{k+1}), y^k, c^k) - y^k \| \, .$$

   (b) *Multiplier update: Define* $y^{k+1} := \nabla_w P(g(x^{k+1}), y^k, c^k)$*.*

■

There was no attempt to exploit any structure arising in the problems[4]. We also did not try to fine tune the parameters of the algorithms to achieve better performance in each problem. The parameters were chosen to be robust, *i.e.* to guarantee convergence to all the problems tested. This was done to avoid conclusions derived from details of implementation.

To solve the unconstrained optimizations we implemented the BFGS trust region with double dog-leg as described in Dennis and Schnabel [4]. We also used the LBFGS-B code from Byrd *et al* [3] which is freely available at the OTC site[5].

Some choices of parameters that must be described are:

1. **Initialization**: The initial primal-dual pair was chosen randomly in $[-2, 2] \times [0, 2]$. This intervals are in the order of magnitude of most solutions.

2. **Stopping criteria**: For both algorithms, we required that the Kuhn-Tucker conditions should be nearly satisfied.

   (a) The relative gradient[6] of the Lagrangian relative to x in the final point, x^* should be smaller than 10^{-3};

   (b) $\max\{| g_i(x^*) |, \lambda_i\} \leqq 10^{-4}$, for all $i = 1, \dots, m$;

   (c) $\| \max\{g(x), 0\} \| \leqq 10^{-4}$;

   (d) $\lambda \geqq 0$.

3. **Update of the penalty parameter c^k**: For both algorithms, we followed the strategy from [2]. Whenever the infeasibility, computed by $\| \max\{0, g(x)\} \|$, did not decrease from a factor of 0.25 then the penalty parameter was multiplied by 7.5.

---

[4] For example, we did not try to use the fact that the majority of the problems chosen are quadratic.

[5] The source code is freely available in the site `http://www-c.mcs.anl.gov/home/otc/`.

[6] See Dennis and Schnabel [4], page 159-161

4. **The σ parameter of the Hybrid method:** We chosen to use $\sigma := 0.15$. It is interesting to say that high values for σ showed worthless, since they increase the number of inner steps required by the method.

5. **The $(\epsilon^k)_{k \in \mathbb{N}}$ of the inexact augmented Lagrangian.** We chosen the harmonic sequence repeated 5 times: $1, 1, 1, 1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \ldots$ This sequence was chosen since it converges slowly to zero and is not summable. The repetition step was chosen trying to guarantee convergence for all the test problems and to minimize the effort of the algorithm.

3.3 Computational environment

All the code was written in Fortran 90, except the LBFGS-B that was in FORTRAN 77. The tests were done using an ultra-SPARC from SUN, with processors at 167 MHz. The operating system is Solaris 2.5.1. We have ran the programs as normal users, in a Sunday, to minimize variation in the load of the processor. The compilers were "WorkShop Compilers 4.2 10/22/96 Fortran 90 1.2" and "WorkShop Compilers 4.2 30 Oct 1996 FORTRAN 77 4.2". Finally we have used the flag "-fast" to optimize the code. Each problem was solved a thousand times to minimize start up effects.

3.4 Results

First, we must stress that both methods were able to solve all the tests, with the above parameters. The Hybrid method presented little more robust behavior than the augmented Lagrangian, apparently due to the addition of the quadratic term in the function being minimized at each iteration.

To compare the methods, we use as performance indicators the total number of function evaluations (objective function, constraints and their gradients) and the processing user time[7].

Since the results of each method in all the tests problems give a huge table, it is preferable to present here only tables with the mean behavior in the test set[8] Of course, anyone interested may contact the authors and receive a copy of the full tables and the programs here quoted[9].

We present two tables, the first showing a comparison between the performance of the inexact augmented Lagrangian, called Bert82, and the Hybrid method using the BFGS trust region dog-leg solver. The second presents the comparison of both methods using the LBFGS-B solver.

In order to obtain the mean behavior, we computed, for each test, the ratio between the performance result of the Bert82 and the Hybrid method. Then we compute the geometric mean of this ratios to get the mean behavior in the test set. So the first entry in the table 1 can be read "The Bert82 implementation performed, in this test set, 75.2% of the total of functions evaluations performed by the Hybrid implementation". In another way: "The Bert82 implementation performs 24.8% less functions evaluations than the Hybrid implementation to solve the problems in this test set". This explain why the Hybrid line is constantly 1.0.

---

[7] Time effectively used, deducing from total time from the other users and system times

[8] As example, the full table for the trust region dog-leg solver is presented in appendix.

[9] Please, send an e-mail to Paulo José da Silva e Silva (`rsilva@ime.usp.br`).

Alg.	Func.	Grad.	Time	Inner
Bert82	0.752	0.727	0.703	0.734
Hybrid	1.000	1.000	1.000	1.000

Table 1: Mean comparison between the Bert82 and Hybrid, using the BFGS trust region dog-leg unconstrained minimizer.

Alg.	Func.	Grad.	Time	Inner
Bert82	0.664	0.652	0.565	0.757
Hybrid	1.000	1.000	1.000	1.000

Table 2: Mean comparison between the Bert82 and Hybrid, using the LBFGS-B from Byrd *et all.*

Finally, since the algorithms always compute the objective function and the constraints (and their derivatives) together, only one column for function evaluations (Func.) and one column for gradients and Jacobian (Grad.) is presented. The column "Inner" is relative to the number of inner iterations required to accept the solution.

4 Conclusions

In this paper we have introduced a method for constrained convex optimization that requires "fixed" precision at each step. As far as the authors are concerned this is the first theoretical result in such direction. The Hybrid method was successful in solving the test problems, showing its feasibility as an constrained optimization method.

Although, the small computational test performed suggests that this method, naively implemented, can not compete with its classical counterpart, the augmented Lagrangian, to solve well behaved problems. The main reason for this drawback is the increased number of inner steps. This can be explained, probably, by the addition in the proximal augmented Lagrangian of a term that penalizes moving away from the last primal point.

We also call attention to the fact that the theoretical result of smaller precision did not imply in less work at each inner step. There are two possible reasons for this fact: first, the projection seems to decrease the connection between the unconstrained problems arising in to consecutive inner steps. Then we do not have such a good starter to the next minimization. On the other hand, it is well known that, in practice, the augmented Lagrangian converges even with small fixed precision as stated in [2]. That is the reason that permitted us to choose the sequence $(\epsilon^k)_{k\in\mathbb{N}}$ converging slowly to zero minimizing the effect of the fixed precision in the Hybrid algorithm.

References

[1] D. Bertsekas. *Constrained Optimization and Lagrange Multipliers.* Academic Press, New York, 1982.

[2] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1995.

[3] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ci You Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[4] D. E.. Dennis Jr. and Robert B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, volume 16 of *Classics in Applied Mathematics 16*. SIAM, 1996.

[5] Willi Hock and Klaus Schittkowski. *Test examples for nonlinear programming codes*, volume 187 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Berlin, 1981.

[6] Leon S. Lasdon. An efficient algorithm for minimizing barrier and penalty functions. *Mathematical Programming*, 2(1):65–106, 1972.

[7] B. Lemaire. The proximal algorithm. In *International Series of Numerical Mathematics (J.P. Penot, editor*, pages 73–87. Birkhauser, Basel, 1989.

[8] B. Martinet. Determination approché d'un point fixe d'une application pseudo-contractante. *C.R. Acad. Sci. Paris*, 274A:163–165, 1972.

[9] J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[10] R. T. Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1:97–116, 1976.

[11] R. T Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:887–898, August 1976.

[12] Klaus Schittkowski. *More test examples for nonlinear programming codes*, volume 282 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Berlin, 1987.

[13] M. Solodov and B. Svaiter. A hybrid projection-proximal point algorithm. *Jounal of Convex Analysis - to appear*, 1997.

A Full Table using BFGS trust region dog-leg solver

Here we present the full performance table of both algorithms using as unconstrained solver the BFGS trust region dog-leg. Remember that the numbers are results of a thousand solutions of the same problem with different randomly chosen starting points.

The columns are organized as follows: "Stat." - Status of the minimization, S if all problems were solved successfully, F otherwise; "Func." - Total number of times that is required to evaluate the objective functions and the constraints; "Grad." - Total number that is required to evaluate the gradient of the objective function and the Jacobian of the constraints; "Time" - user time spent; "Inner" - Total number of inner iterations.

Prob.	Alg.	Stat.	Func.	Grad.	Time	Inner
S21	Bert82	S	45051	21961	6.83E+00	3356
S21	Hybrid	S	66727	41241	1.19E+01	8218
S215	Bert82	S	108438	47068	1.13E+01	14366
S215	Hybrid	S	122255	53616	1.39E+01	16199
S218	Bert82	S	65184	31665	7.75E+00	8001
S218	Hybrid	S	89432	39803	1.00E+01	12535
S224	Bert82	S	57583	38169	1.09E+01	8607
S224	Hybrid	S	80501	47040	1.47E+01	6921
S262	Bert82	S	233128	116627	4.49E+01	11900
S262	Hybrid	S	144146	103908	4.06E+01	11469
S268	Bert82	S	133527	80586	3.39E+01	4455
S268	Hybrid	S	366292	223798	1.01E+02	16648
S28	Bert82	S	192468	80082	2.54E+01	4570
S28	Hybrid	S	276347	134290	4.26E+01	17361
S284	Bert82	S	111481	97308	7.77E+01	6136
S284	Hybrid	S	237252	186295	1.59E+02	10945
S315	Bert82	S	50439	31946	8.01E+00	7381
S315	Hybrid	S	68500	46026	1.10E+01	12039
S35	Bert82	S	66233	44983	1.37E+01	8029
S35	Hybrid	S	98216	65214	1.98E+01	14466
S384	Bert82	S	159377	120114	9.36E+01	12614
S384	Hybrid	S	323945	263956	2.25E+02	13344
S51	Bert82	S	248894	120081	4.67E+01	20498
S51	Hybrid	S	261081	158608	6.91E+01	12999
S76	Bert82	S	74651	63188	2.14E+01	13371
S76	Hybrid	S	108611	88017	2.92E+01	18407
Lasdon1	Bert82	S	68306	48788	1.12E+01	14211
Lasdon1	Hybrid	S	65976	46557	1.19E+01	12932
Lasdon3	Bert82	S	133845	118125	3.77E+01	21989
Lasdon3	Hybrid	S	109396	85703	2.99E+01	12166
p4	Bert82	S	55523	47028	1.39E+01	8885
p4	Hybrid	S	65038	50532	1.40E+01	10877

Table 3: Comparison between Bert82 and Hybrid using BFGS trust dog-leg solver.

Linear Optimal Cost Cumulant Control: A k-Cumulant Problem Class[1]

Khanh D. Pham	Stanley R. Liberty	Michael K. Sain
Dept. of Electrical Eng.	Office of the Provost and VPAA	Dept. of Electrical Eng.
University of Nebraska-Lincoln	Bradley University	University of Notre Dame
Lincoln, Nebraska 68588	Peoria, Illinois 61625	Notre Dame, Indiana 46556

Abstract

The simplest linear stochastic regulator control problem is revisited from a new perspective. An optimal control problem is formulated in which the objective is minimization of a finite, linear combination of cumulants of integral quadratic cost over linear, memoryless, full-state-feedback control laws. The standard linear-dynamical system constraint on a finite time interval with additive Gaussian-white noise and a non-random initial state underlies the problem formulation. Because of the linearity assumptions in the problem statement, it can be formulated as a non-stochastic optimization problem utilizing equations for cost cumulants developed in the 1970s. Since this problem formulation is parameterized both by the number of cumulants and by the scalar coefficients in the linear combination, it defines a very general Linear-Quadratic-Gaussian (LQG) problem class. The special case where only the first cost cumulant is minimized is, of course, the well known minimum mean LQG problem. The general solution of the k-cumulant problem is presented herein. For a particular k-cumulant problem the solution is determined by solving a coupled set of k, matrix, Riccati-type equations. The issue of existence of solution to the optimization problem becomes that of existence of solutions to the Riccati-type equations. However conditions ensuring existence of solutions to these equations are not known. In fact, for values of linear combination coefficients outside certain finite ranges, the equations exhibit finite escape time behavior. On the other hand, for limited ranges of the combination coefficient values, the equations are well behaved and even yield steady-state (stationary) solutions. An example is included illustrating some closed-loop system properties for a 3-cumulant controller.

1 Introduction

In 1964 Sain first reported research results on optimal stochastic control problems where the optimization criteria involved both the variance and the mean of the cost functionals (performance measures) used in the problem statements [8, 9, 10, 11]. Since then, Sain and colleagues have investigated a variety of issues within the problem class referred to as minimum-cost-variance (MCV) control - a problem class where one seeks control laws which minimize the cost variance while enforcing a cost mean constraint [15, 17, 18]. In 1994 Won [16] solved the full-state-feedback MCV problem, and showed that the optimal control law could be calculated by solving a coupled pair of Riccati-type equations. More recently the output feedback MCV problem has yielded some results [6, 7].

Sain, in 1967, also showed interest in cost statistics beyond the mean and variance and published results on cost moment recursions [12]. The matter of finding a complete statistical description of a linear system's performance in a noisy environment was investigated by Liberty [3]. In that work Liberty observed that for a linear system with

---

[1] Funded by the Frank M. Freimann Chair in Electrical Engineering, University of Notre Dame, and by the National Science Foundation under Grants CMS 93-01584, CMS 95-00301, and CMS 95-28083.

Gaussian-white input noise and Gaussian initial state, all cumulants of any finite-time integral-quadratic-form (IQF) in the system state were quadratic-affine in the initial state mean. Prior to that it was only known that the mean and the variance (which happen to be the first and second cumulants) of such IQFs had this form. Subsequently Liberty and Hartwig [4] showed that these cumulants could be calculated by solving a set of coupled matrix differential equations. Although at that time these equations were not utilized in the formulation of an optimal control problem, Liberty [5] did develop ad hoc feedback control laws from the equations associated with the first two cumulants and observed in examples that these feedback control laws performed as one might expect MCV control laws to perform. Indeed, the ad hoc Riccati-type equations displayed in [5] are essentially the same as those developed by Won in [16].

In view of the recent MCV research successes mentioned above, and by the closed-loop properties exhibited by MCV control laws in applications [17], it is desirable to formulate and solve higher order cost cumulant control problems. However formulating even a third cost cumulant problem along the lines of that done in [16] is a formidable task. Thus we returned to the Liberty and Hartwig equations [4] to see if they could be used to formulate a linear optimal cost cumulant control problem. The subsequent sections of this paper contain the results of this investigation.

2 Preliminaries

Consider a linear dynamical system modeled on the time interval $T = [t_0,\, t_f]$ by the stochastic differential equation

$$dx(t) = F(t)x(t)dt + G(t)dw(t) \tag{1}$$

where the system state $x(t) \in \mathcal{R}^n$ and the noise $w(t) \in \mathcal{R}^p$ is a stationary Wiener process with correlation of increments $E\{[w(t) - w(\sigma)][w(t) - w(\sigma)]^T\} = W|t - \sigma|$. Assume that the initial state $x(t_0)$ is non-random. Associate with (1) a functional

$$J = x^T(t_f)Q_f x(t_f) + \int_{t_0}^{t_f} [x^T(t)N(t)x(t)]dt \tag{2}$$

where Q_f and $N(t)$ are symmetric and positive semi-definite. Also, assume that $F(\cdot)$, $G(\cdot)$, and $N(\cdot)$ are continuous on T.

For the above system and associated functional J, from [4] the equations which determine the k^{th} cumulant of J are

$$\kappa_k = x_0^T H(t_0,\, k)x_0 + TrD(t_0,\, k), \tag{3}$$

where $Tr\{\cdot\}$ denotes the trace of the enclosed matrix and $H(t_0,\, i)$ and $D(t_0,\, i)]$ are solutions at $t = t_0$ of the following coupled matrix differential equations:

$$\begin{aligned}
\frac{d}{dt}H(t,\, 1) &= -F^T(t)H(t,\, 1) - H(t,\, 1)F(t) - N(t) && (4)\\
\frac{d}{dt}H(t,\, i) &= -F^T(t)H(t,\, i) - H(t,\, i)F(t) && (5)\\
&\quad - \sum_{j=1}^{i-1} \frac{2i!}{j!(i-j)!} H(t, j)G(t)WG^T(t)H(t, i-j), \qquad i \geq 2 &&\\
\frac{d}{dt}Tr\{D(t,\, i)\} &= -Tr\{H(t,\, i)G(t)WG^T(t)\}, \qquad i \geq 1 && (6)
\end{aligned}$$

with terminal conditions $H(t_f,\, 1) = Q_f$, $H(t,\, i) = 0$ for $i \geq 2$, and $D(t,\, i) = 0$ for $i \geq 1$.

3 The k-Cumulant Problem Statement

Now consider a linear dynamical system with control, $u(t) \in \mathcal{R}^m$, and state $x_c(t) \in \mathcal{R}^n$ modeled on T by

$$\dot{x}_c(t) = (A(t)x_c(t) + B(t)u(t))dt + G(t)dw(t) \tag{7}$$

where $A(\cdot)$, and $B(\cdot)$ are continuous on T and the control, $u(\cdot)$ is a linear, memoryless, full-state feedback control given by $u(t) = K(t)x_c(t)$ where $K(\cdot)$ is any $m \times n$ matrix feedback gain which is in the class $\mathcal{K}_k$ of admissible feedback gains defined in Definition 1 below. The process $w(\cdot)$ and the coefficient $G(\cdot)$ for the system described by (7) are the same as in (1). Also, we assume that both (1) and (7) have the same known initial condition, i.e., $x_c(t_0) = x(t_0)$. Attach to (7) the cost functional

$$J_c = x_c^T(t_f)Q_f x_c(t_f) + \int_{t_0}^{t_f} [x_c^T(\tau)Q(\tau)x_c(\tau) + u^T(\tau)R(\tau)u(\tau)]d\tau \tag{8}$$

where $Q(\tau)$ and $R(\tau)$ are symmetric and continuous on T with $Q(\tau)$ positive semidefinite and $R(\tau)$ positive definite. Also Q_f is the same as in (2).

If $F(\cdot)$ and $N(\cdot)$ in (1) and (2) are given by $F(t) = A(t) + B(t)K(t)$ and $N(t) = Q(t) + K^T(t)R(t)K(t)$, then for every feedback gain $K(\cdot)$ in $\mathcal{K}_k$ then $x(\cdot) = x_c(\cdot)$ and $J = J_c$. Therefore, for each admissible gain $K(\cdot)$ the k^{th} cumulant of J_c is given by (3) where

$$\begin{aligned} \frac{d}{dt}H(t,1) &= -A^T(t)H(t,1) - H(t,1)A(t) - Q(t) \\ &\quad -K^T(t)B^T(t)H(t,1) - H(t,1)B(t)K(t) - K^T(t)R(t)K(t) \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{d}{dt}H(t,i) &= -A^T(t)H(t,i) - K^T(t)B^T(t)H(t,i) - H(t,i)B(t)K(t) \\ &\quad -H(t,i)A(t) - \sum_{j=1}^{i-1} \frac{2i!}{j!(i-j)!} H(t,j)G(t)WG^T(t)H(t,i-j), \qquad i \geq 2 \end{aligned} \tag{10}$$

$$\frac{d}{dt}Tr\{D(t,i)\} = -Tr\{H(t,i)G(t)WG^T(t)\}, \qquad i \geq 1 \tag{11}$$

with terminal conditions $H(t_f, 1) = Q_f$, $H(t,i) = 0$ for $i \geq 2$, and $D(t,i) = 0$ for $i \geq 1$.

Now the class, $\mathcal{K}_k$ of admissible feedback gains can be defined as follows:

Definition 1. *Let $\overline{K}$ be a compact subset of the vector space of all $m \times n$ real matrices. ($\overline{K}$ is the allowable set of gain values.) Let $\mathcal{K}_k$ be the set of all $m \times n$ matrix functions which are continuous on T with values $K(t) \in \overline{K}$, and which yield existence of solutions on T to the k coupled, matrix differential equations (9) through (11) with the terminal conditions specified above.*

Form the partitioned matrix $Y^T(t) = [\,H^T(t,1) \quad D^T(t,1) \;, \ldots \quad H^T(t,k) \quad D^T(t,k)\,]$ and let $y(t) = vec[Y(t)]$ where the operation $vec[Y(t)]$ produces the $2kn^2 \times 1$ partitioned matrix with the columns of $Y(t)$ as entries. Also notate the relationship among $\dot{Y}$, Y, and K as described in (9) through (11) with the specified terminal conditions by

$$\frac{d}{dt}Y(t) = f(t, Y(t), K(t)), \qquad Y(t_f) = Y_f. \tag{12}$$

Definition 2. *Let $\phi_0\big(t_0, Y(t_0); x(t_0); \mu_1, \ldots, \mu_k\big) = \sum_{i=1}^{k} \mu_i \kappa_i$ be the performance index, where each μ_i is a non-negative real scalar constant.*

Without loss of generality, set $\mu_1 = 1$. The k-cumulant optimization problem can now be seen as the minimization $\min\limits_{\mathrm{K}(\cdot) \in \mathcal{K}_\mathrm{k}} \phi_0\big(t_0, Y(t_0); x(t_0); \mu_2, \ldots, \mu_k\big)$ subject to (12).

Although this optimization criterion represents a competition among cumulant values, the ultimate objective herein is to introduce parametric freedom in the class of feedback control laws which will result from the problem solution. Hopefully this parametric freedom can be exploited to achieve desirable closed-loop system properties. Since the cumulant values are constrained by the nature of the distribution of the random variable, J (generalized chi-square), one should expect that solutions to the optimization problem will exist only for limited ranges of the values of the combination coefficents.

This optimization problem is in "Mayer form" and can be solved by applying the Mayer form verification theorem of dynamic programming given in Fleming and Rishel [2]. Note that the Hamilton-Jacobi-Bellman equation (HJB) for the Mayer form problem is

$$\min_{K \in \overline{K}} \left\{ \frac{\partial}{\partial t} V(t, Y) + \frac{\partial}{\partial y} V(t, Y) \cdot vec\big[f(t, Y, K)\big] \right\} = 0 \tag{13}$$

where $V(\cdot, \cdot)$ is a real valued functional. The verification theorem stated in the notation used here is:

Theorem 1. *For a fixed value of k, let $V(t, Y)$ be a continuously differentiable solution of the HJB equation (13) which satisfies the boundary condition $V(t_0, Y) = \phi_0(t_0, Y(t_0))$ Let $K(\cdot)$ be any admissible control gain and let $Y(\cdot)$ be the corresponding solution of (12). Then $V(t, Y(t))$ is a non-increasing function of t (since (12) has a specified final condition). If $K^*(\cdot)$ is a control gain with corresponding solution, $Y^*(\cdot)$, of (12) such that for $t \in [t_0, t_f]$*

$$\frac{\partial}{\partial t} V(t, Y^*(t)) + \frac{\partial}{\partial y} V(t, Y^*(t)) \cdot vec\big[f(t, Y^*(t), K^*(t))\big] = 0 \tag{14}$$

then $K^*(\cdot)$ minimizes the performance index and

$$V(t, Y) = inf\big\{\phi_0(t_0, Y(t_0); x(t_0); \mu_2, \ldots, \mu_k)\big\} \tag{15}$$

where the infimum is taken over all admissible feedback gains on the sub-interval $[t_0, t] \subset [t_0, t_f]$ with terminal condition $Y(t)$. The functional $V(\cdot, \cdot)$ which satisfies (15) is called the value function.

4 k-Cumulant Control Solution

To solve the k-cumulant optimization problem stated in Section 3, guess a solution to (13) of the form,

$$V(t, Y) = Tr\Big\{ M(t)\big[H_1 + \mu_2 H_2 + \ldots + \mu_k H_k\big] + \big[D_1 + \mu_2 D_2 + \ldots + \mu_k D_k\big] \Big\} \tag{16}$$

where $M(\cdot)$ is a continuously differentiable matrix function with $M(t_0) = x_0 x_0^T$ and the variable Y is given by $Y^T = [\, H_1^T \quad D_1^T \quad \ldots \quad H_k^T \quad D_k^T \,]$ where each H_i and each D_i is an $n \times n$ matrix variable. It is necessary to verify that the guess in (16) satisfies the HJB equation, the boundary condition

$$\begin{aligned} V(t_0, Y(t_0)) \;&=\; Tr\Big\{ M(t_0)\big[H_1(t_0) + \mu_2 H_2(t_0) + \ldots + \mu_k H_k(t_0)\big] \\ &\quad + \big[D_1(t_0) + \mu_2 D_2(t_0) + \ldots + \mu_k D_k(t_0)\big] \Big\}, \end{aligned} \tag{17}$$

and that it satisfies (14) along the $Y(t)$ trajectory produced by the $K(\cdot)$ resulting from the minimization in (13). Clearly, for the initial condition given for $M(t_0)$, (17) is satisfied. Under the feedback gain the H equations (9) through (10) become Riccati-type equations:

$$\begin{aligned}
\dot{H}(t,1) &= -A^T(t)H(t,1) - H(t,1)A(t) + H(t,1)B(t)R^{-1}(t)B^T(t)H(t,1) \qquad (18)\\
&\quad -Q(t) - \sum_{i=2}^{k} \mu_i H(t,i)B(t)R^{-1}(t)B^T(t)\sum_{j=2}^{k} \mu_j H(t,j)\\
\dot{H}(t,i) &= -A^T(t)H(t,i) - H(t,i)A(t) + H(t,1)B(t)R^{-1}(t)B^T(t)H(t,i) \qquad (19)\\
&\quad +\sum_{j=2}^{i} \mu_j H(t,j)B(t)R^{-1}(t)B^T(t)H(t,i) + H(t,i)B(t)R^{-1}(t)B^T(t)H(t,1)\\
&\quad +H(t,i)B(t)R^{-1}(t)B^T(t)\sum_{j=2}^{i} \mu_j H(t,j)\\
&\quad -\sum_{j=1}^{i-1} \frac{2i!}{j!(i-j)!} H(t,j)G(t)WG^T(t)H(t,i-j). \qquad i \geq 2.
\end{aligned}$$

Thus, if the Riccati-type equations have positive definite solutions, then the existence of an $M(\cdot)$ satisfying

$$\begin{aligned}
\dot{M}(t) &= -M(t)\Big[-A^T(t)H(t,1) - H(t,1)A(t) - Q(t) \qquad (20)\\
&\quad +H(t,1)B(t)R^{-1}(t)B^T(t)H(t,1) - \sum_{i=2}^{k} \mu_i H(t,i)B(t)R^{-1}(t)B^T(t)\sum_{j=2}^{k} \mu_j H(t,j)\\
&\quad +\sum_{i=2}^{k} \mu_i\Big(-A^T(t)H(t,i) - H(t,i)A(t) + H(t,1)B(t)R^{-1}(t)B^T(t)H(t,i)\\
&\quad +\sum_{l=2}^{k} \mu_l H(t,l)B(t)R^{-1}(t)B^T(t)H(t,i) + H(t,i)B(t)R^{-1}(t)B^T(t)H(t,1)\\
&\quad +H(t,i)B(t)R^{-1}(t)B^T(t)\sum_{l=2}^{k} \mu_l H(t,l)\\
&\quad -\sum_{j=1}^{i-1} \frac{2i!}{j!(i-j)!} H(t,j)G(t)WG^T(t)H(t,i-j)\Big)\Big]\Big[H(t,1) + \sum_{j=2}^{k} \mu_j H(t,j)\Big]^{-1}\\
&\quad +\Big[H(t,1) + \sum_{i=2}^{k} \mu_i H(t,i)\Big]G(t)WG^T(t)\Big]\Big[H(t,1) + \sum_{j=2}^{k} \mu_j H(t,j)\Big]^{-1}
\end{aligned}$$

is standard. Enforcing the terminal value conditions, $H(t_f,1) = Q_f$, $H(t_f,i) = 0$ for $i \geq 2$ uniquely specifies $M(\cdot)$ and yields a value function which satisfies (14) of the verification theorem. Thus the optimal feedback gain which minimizes the performance index follows:

$$K(t) = -R^{-1}(t)B^T(t)\Big[H(t,1) + \sum_{i=2}^{k} \mu_i H(t,i)\Big]. \qquad (21)$$

Note that the existence of a solution to the k-cumulant optimization problem hinges on the existence of solutions to equations (18) through (19) on T. At this time, no explicit conditions are known that will ensure the existence of solutions on T to equations (18) through (19) for non-zero values of μ_2 through μ_k. However, it can readily be seen, from the series expressions for the cumulants of J contained in [3] and [4] and the fact that all of the cumulants are non-negative, that if solutions exist to (18) through (19), then they are positive semidefinite on T. Furthermore, if $H(\cdot,1)$ is positive definite on $[t_0,t_f)$

then $H(\cdot,2),\ldots,H(\cdot,k)$ are also positive definite on $[t_0,t_f)$. It is also not known at this time if positive definiteness of solutions to the Riccati-type equations ensures positive semi-definiteness of $[M(t)+M^T(t)]$ on T, but that is the conjecture made here.

5 Example: Single Degree of Freedom (SDOF) Structure with a 3-Cumulant Controller

Next we show how the additional parametric freedom of the k-cumulant problem class influences some closed-loop system properties. For the $k=3$ case the optimization criterion is

$$\min_{K\in\bar{\mathcal{K}}_k} \phi_0\big(t_0,\, Y(t_0);\, x(t_0);\, \mu_2,\, \mu_3\big) = \big\{\kappa_1+\mu_2\kappa_2+\mu_3\kappa_3\big\},$$

and the equations (18) through (19) with their terminal conditions are:

$$\begin{aligned}
\dot{H}(t,1) &= -A^T(t)H(t,1)-H(t,1)A(t)-Q(t)+H(t,1)B(t)R^{-1}(t)B^T(t)H(t,1) \qquad (22)\\
&\quad -(\mu_2)^2H(t,2)B(t)R^{-1}(t)B^T(t)H(t,2)-\mu_2\mu_3H(t,2)B(t)R^{-1}(t)B^T(t)H(t,3)\\
&\quad -\mu_2\mu_3H(t,3)BR^{-1}(t)B^TH(t,2)-(\mu_3)^2H(t,3)BR^{-1}(t)B^TH(t,3),\\
\dot{H}(t,2) &= -A^T(t)H(t,2)-H(t,2)A(t)+2\mu_2H(t,2)B(t)R^{-1}(t)B^T(t)H(t,2) \qquad (23)\\
&\quad +H(t,2)B(t)R^{-1}(t)B^T(t)H(t,1)+H(t,1)B(t)R^{-1}(t)B^T(t)H(t,2)\\
&\quad +\mu_3H(t,2)B(t)R^{-1}(t)B^T(t)H(t,3)+\mu_3H(t,3)B(t)R^{-1}(t)B^T(t)H(t,2)\\
&\quad -4H(t,1)G(t)WG^T(t)H(t,1),\\
\dot{H}(t,3) &= -A^T(t)H(t,3)-H(t,3)A(t)+H(t,3)B(t)R^{-1}(t)B^T(t)H(t,1) \qquad (24)\\
&\quad +H(t,1)B(t)R^{-1}(t)B^T(t)H(t,3)+\mu_2H(t,3)B(t)R^{-1}(t)B^T(t)H(t,2)\\
&\quad +\mu_2H(t,2)B(t)R^{-1}(t)B^T(t)H(t,3)+2\mu_3H(t,3)B(t)R^{-1}(t)B^T(t)H(t,3)\\
&\quad -6H(t,1)G(t)WG^T(t)H(t,2)-6H(t,2)G(t)WG^T(t)H(t,1),
\end{aligned}$$

with terminal conditions $H(t_f,1)=Q_f$, $H(t_f,2)=0$, and $H(t_f,3)=0$. From (21), the optimal gain is

$$K(t) = -R^{-1}(t)B^T(t)[H(t,1)+\mu_2H(t,2)+\mu_3H(t,3)]. \qquad (25)$$

The design procedure is to select values of Q and R in the integral quadratic cost functional, as well as values for μ_2 and μ_3 such that closed-loop system has favorable properties.

The existence of solutions to equations (22) through (24) can not be ensured for arbitrary values of μ_2 and μ_3. Indeed, the equations exhibit finite escape time for values of μ_2 or μ_3 lying outside limited ranges. The lower limits on the values are both zero. The upper limits on these ranges are found by trial and error and depend on the specific dynamics of the system under consideration.

The optimal k-cumulant controllers are time-varying. However, it is desirable in practice to implement time-invariant controllers which approximate the behavior of the optimal time-varying controllers. This can be done if the solutions to equations (22) through (24) reach steady state values in a time interval that is much smaller than T. This is the case in the example studied here. The time-invariant feedback gains corresponding to steady-state solutions to (22) through (24) are used for examining closed-loop system properties. In this case the degradation of the performance index by utilizing sub-optimal controllers is calculated and shown to be very slight.

A useful measure of closed-loop system performance is the standard deviation (RMS value) of a component of the closed-loop system's state. For a time-invariant system with stationary noise and a time-invariant controller, the standard deviations (RMS values) of all sytem states can be calculated by finding the solution to the linear algebraic equation describing the steady-state covariance, Σ, of the closed-loop system state vector,

$$A_{CL}\Sigma_x + \Sigma_x A_{CL}^T + GWG^T = 0 \tag{26}$$

where A_{CL} is the closed-loop state matrix.

The 3-cumulant controller is now applied to the problem of controlling vibration effects in a lightly damped single degree of freedom structure with tendon control during a seismic event. The system model is given in [1] as:

$$dx(t) = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix} x(t)dt + \begin{bmatrix} 0 \\ -\frac{4k_c cos\alpha}{m} \end{bmatrix} u(t)dt + \begin{bmatrix} 0 \\ -1 \end{bmatrix} dw(t), \tag{27}$$

where $x_1(t)$ is floor position and $x_2(t)$ is floor velocity. The system constants are: $m = 14.69lb - sec^2/in$, $c = 9.020lb - sec/in$, $k = 7934lb/in$, $k_c = 2124lb/in$, $\alpha = 36$ degrees and the noise covariance is $W = 1.00 \times 2\pi\ in^2/sec^3$. For this example, we let $x_0 = [0.2\ 0]^T$, and select $R = 2124$ and $Q = diag.(7934, 0)$. With terminal penalty weighting matrix Q_f equal to zero and with values of μ_2 and μ_3 in the interval $[0, 9]$.

6 Computation Results

In Figure 1, the optimal cumulant values resulting from different values for μ_2 and μ_3 are shown. Observe that the $k = 3$ case allows more flexibility in affecting the cumulant values than the case $k = 2$ (which corresponds to $\mu_3 = 0$). In Figure 2 standard deviations (RMS values) of floor displacement and floor velocity are shown for different values of μ_2 and μ_3. Note that for the $k = 3$ problem solution the standard deviation of the floor displacement and velocity can both be made smaller than in the cases $k = 1$ ($\mu_2 = \mu_3 = 0$) and $k = 2$ ($\mu_3 = 0$). This improvement in state regulation is accomplished by control action with standard deviation greater than that of the lower order problem solutions.

References

[1] L. L. Chung, A. M. Reinhorn and T. T. Soong, "Experiments on Active Control of Seismic Structures," *Journal of Engineering Mechanics,* Vol. 114, pp. 241-256, 1988.

[2] Wendell H. Fleming and Raymond W. Rishel, *Deterministic and Stochastic Optimal Control,* Springer-Verlag, pp. 87-88, 1975.

[3] S. R. Liberty, "Characteristic Functions of LQG Control," *Ph.D. Dissertation,* Department of Electrical Engineering, University of Notre Dame, August 1971.

[4] Stanley R. Liberty and Ronald C. Hartwig, "On the Essential Quadratic Nature of LQG Control-Performance Measure Cumulants," *Information and Control,* Vol. 32, No. 3, pp. 276-305, November 1976.

[5] Stanley R. Liberty, "Performance Analysis of Stochastic Linear Control Systems: A New Viewpoint," *Proceedings International Forum on Alternatives for Linear Multivariable Control,* pp. 79-86, October 1977.

[6] S. R. Liberty, Peter Quast, and M. K. Sain, "Control Access to Performance Measure Statistics: A Second Moment Case," *Proceedings Allerton Conference on Communication, Control, and Computing,* pp. 933-942 , 1997.

[7] Peter Quast, "Output Feedback Minimum Cost Variance Control Schemes with Application To Seismic Protection of Civil Engineering Structures," *Ph.D. Dissertation,* Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, September 1997.

[8] M. K. Sain, "On Minimal-Variance Control of Linear Systems with Quadratic Loss," *Ph.D. Dissertation,* Department of Electrical Engineering and Coordinated Science Laboratory, University of Illinois, Urbana, Illinois, January 1965.

[9] M. K. Sain, "Relative Costs of Mean and Variance Control for a Class of Linear, Noisy Systems," *Proceedings Allerton Conference on Circuit and System Theory,* pp. 121-129, 1965.

[10] M. K. Sain, "Control of Linear Systems According to the Minimal Variance Criterion-A New Approach to the Disturbance Problem," *IEEE Transactions on Automatic Control,* AC-11, No. 1, pp. 118-122, January 1966.

[11] M. K. Sain, "A Sufficiency Condition for Minimum Variance Control of Markov Processes," *Proceedings Allerton Conference on Circuit and System Theory,* pp. 593-599, 1966.

[12] M. K. Sain, "Performance Moment Recursions, with Application to Equalizer Control Laws," *Proceedings Allerton Conference on Circuit and System Theory,* pp. 327-336, 1967.

[13] M. K. Sain and S. R. Liberty, "Performance Measure Densities for a Class of LQG Control Systems," *IEEE Transactions on Automatic Control,* AC-16, No. 5, pp. 431-439, October 1971.

[14] M. K. Sain, Chang-Hee Won, and B. F. Spencer, Jr. "Cumulant Minimization and Robust Control," in *Stochastic Theory and Adaptive Control, Lecture Notes in Control and Information Sciences 184,* T. E. Duncan and B. Pasik-Duncan (Eds.), Springer-Verlag, pp. 411-425, 1992.

[15] M. K. Sain, C.-H Won, and B. F. Spencer, "Cumulants in Risk-Sensitive Control: The Full-State-Feedback Cost Variance Case," *Proceedings IEEE Conference on Decision and Control,* New Orleans, pp. 3247-3252, December 1995.

[16] Chang-Hee Won, "Cost Cumulants in Risk-Sensitive and Minimal Cost Variance Control," *Ph.D. Dissertation,* Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, November 1994.

[17] C.-H Won, M. K. Sain, and B. F. Spencer, Jr., "Performance and Stability Characteristics of Risk-Sensitive Controlled Structures under Seismic Disturbances," *Proceedings American Control Conference,* Seattle, Washington, pp. 1926-1930, June 21, 1995.

[18] Chang-Hee Won, Michael K. Sain, and Billie F. Spencer Jr.,"Relationships in Modern Stochastic Control: Risk-Sensitive and Minimal Cost Variance Control," *Proceedings Allerton Conference on Communication, Control, and Computing,* pp. 343-352, 1996.

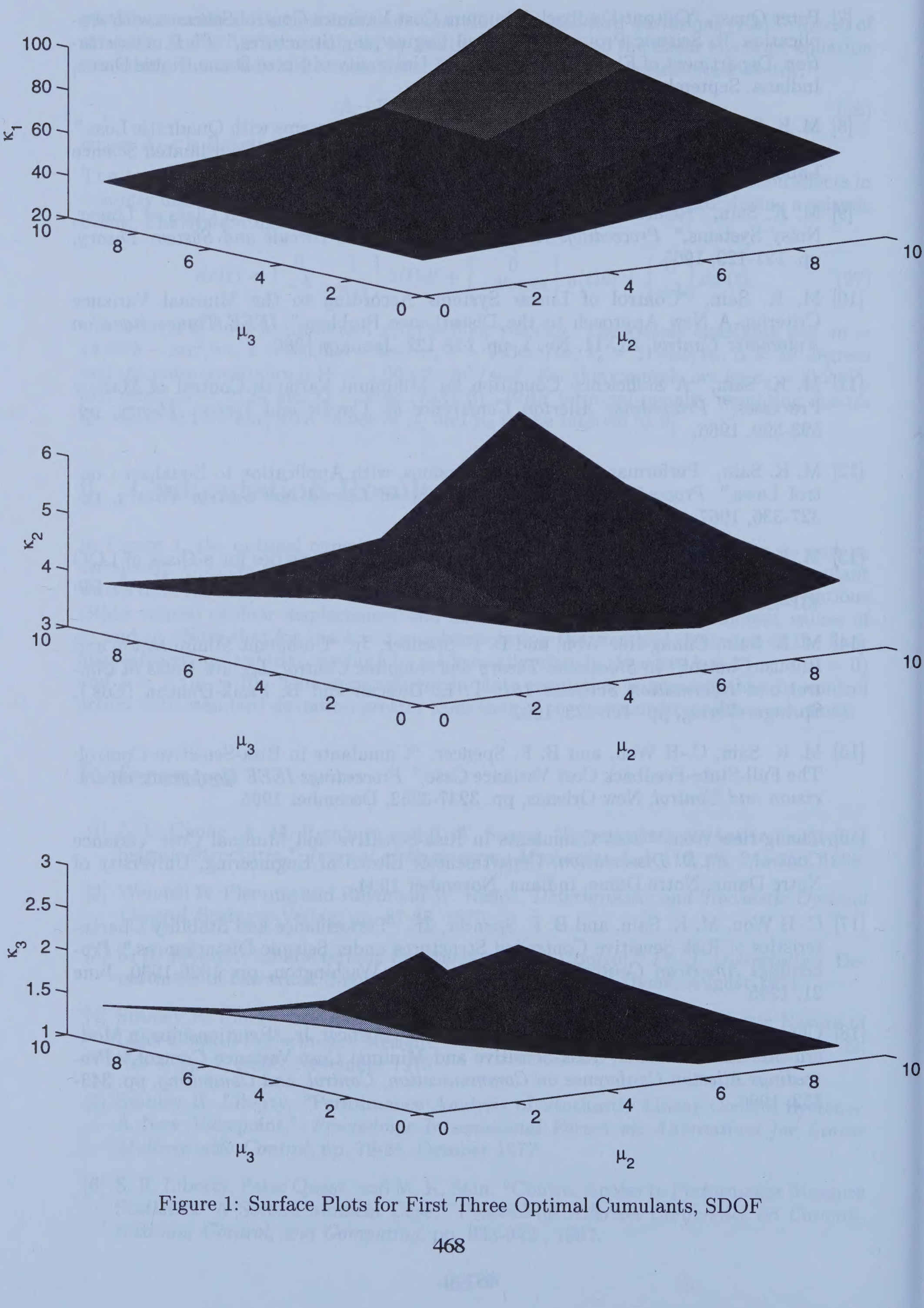


Figure 1: Surface Plots for First Three Optimal Cumulants, SDOF

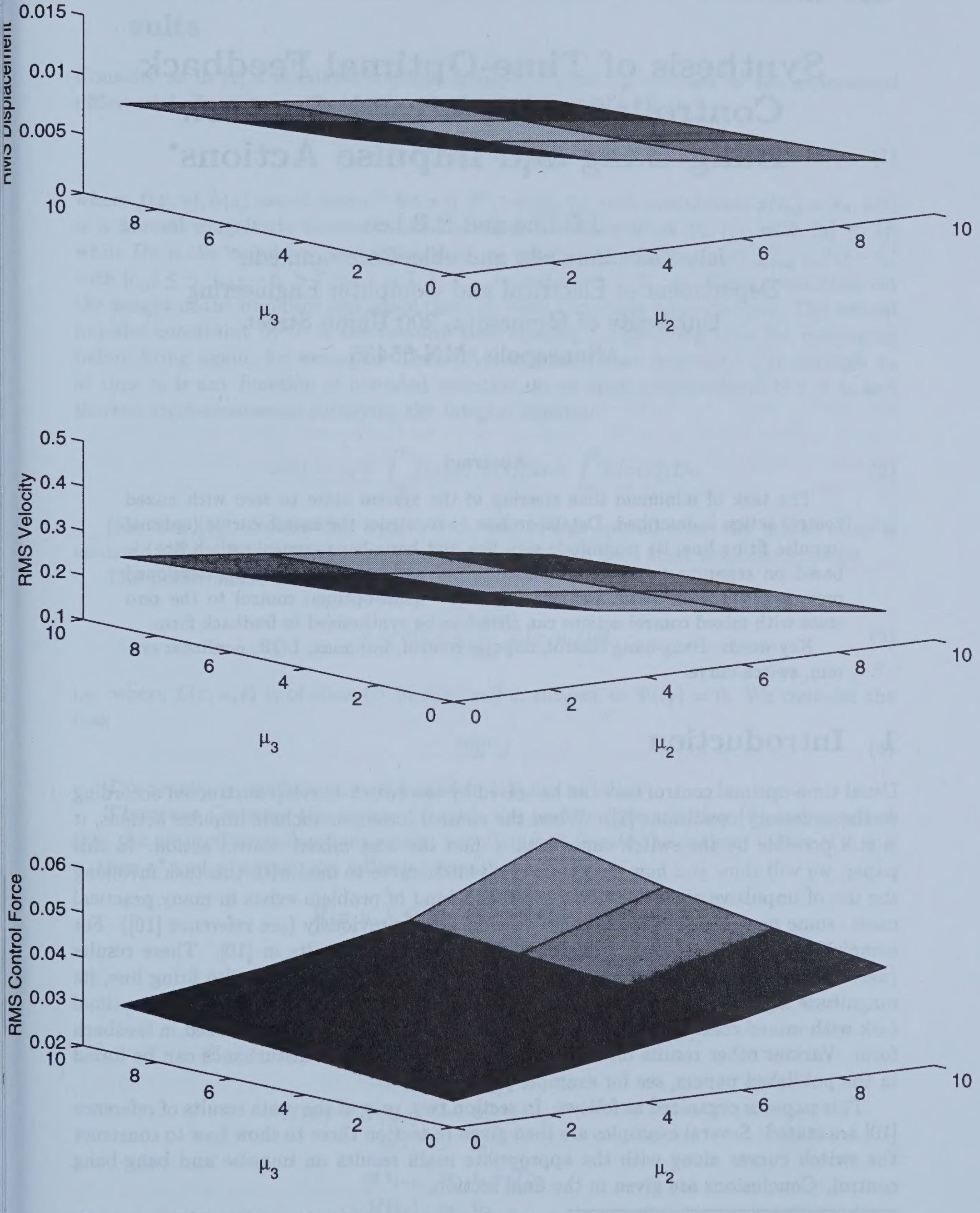


Figure 2: Surface Plots for Displacement, Velocity, and Control Force Standard Deviations, SDOF

Synthesis of Time-Optimal Feedback Controllers with Mixed Control: Bang-Bang and Impulse Actions*

J.C.Luo and E.B.Lee
jcluo@ece.umn.edu and eblee@ece.umn.edu
Department of Electrical and Computer Engineering
University of Minnesota, 200 Union Street
Minneapolis, MN 55455

Abstract

The task of minimum time steering of the system state to zero with mixed control action is described. Details on how to construct the switch curves (optimal impulse firing line, its magnitude sign line and bang-bang control switch line) is based on separate results for impulse control and bang-bang control developed previously for some basic nonlinear systems. Time-optimal control to the zero state with mixed control actions can therefore be synthesized in feedback form.

Key words: Bang-bang control, impulse control, minimax, LQR, nonlinear system, switch-curve.

1 Introduction

Usual time-optimal control task can be solved by the switch-curve (constructed according to the necessary conditions [2]). When the control functions include impulse actions, it is still possible by the switch curve idea to find the best mixed control action. In this paper, we will show you how to construct a switch curve to deal with this task involving the use of impulsive control actions. Since this kind of problem exists in many practical cases. some new necessary conditions were obtained previously (see reference [10]). For completeness, we also briefly describe some of the main results in [10]. These results (necessary conditions) help construct the switch curves (optimal impulse firing line, its magnitude sign line and bang-bang control switch line). In this way, the time-optimal task with mixed control (bang-bang and impulse actions) can be synthesized in feedback form. Various other results on impulsive control or impulsive disturbance can be found in the published papers, see for example [2]-[7].

This paper is organized as follows. In section two, part of the main results of reference [10] are stated. Several examples are then given in section three to show how to construct the switch curves along with the appropriate main results on impulse and bang-bang control. Conclusions are given in the final section.

---

*Research supported by the NSF grant DMS-9308007 and by the Vincentine Hermes-Luh endowment.

2 Task Formulation and Summary of Previous Results

Consider as in [2] a nonlinear impulse dynamic system as modeled by the generalized differential equation system in R^n using generalized derivative D:

$$Dx = f(x,u) + h(x)Dv \tag{1}$$

where $f(x,u), h(x)$ are of class C^1 for $x \in R^n, t \in [t_0, t_f]$ with initial state $x(t_0) = x_0, u(t)$ is a normal magnitude limited control variable (measurable on $[t_0, t_f]$, with $|u| \leq 1$), while Dv is the impulsive control functions we use $h(x)Dv = \sum_{j=1}^{m} h_j(x) \sum_{i=1}^{n} c_{ij}\delta(t - t_i)$ with $|c_{ij}| \leq \gamma$, $t_{i+1} - t_i \geq T$ for $j = 1, 2, 3, ..., m$ and $i = 1, 2, 3, ...n$. These constraints set the weight of the impulses and space them at least T seconds apart in time. The second impulse constraint by T is the impulse time spacing t_i, involving time for recharging before firing again, for example. Then [2] a solution (state response) $x(t)$ through x_0 at time t_0 is any function of bounded variation on an open neighborhood of $t = t_0$ and thereon right-continuous satisfying the integral equation:

$$x(t) = x_0 + \int_{t_0}^{t} f(x(s), u(s))ds + \int_{t_0}^{t} h(x(s))Dv \tag{2}$$

(See also [5]). Notice that $x(t_{0-}) = x_0$ and in this special sense the initial value is assumed. We will formulate an optimal control task for such state response systems.

The performance function to be minimized is

$$J = \phi(t_f) + \int_{t_0}^{t_f} L(x,u,t)dt \tag{3}$$

i.e. where $L(x,u,t)$ is of class C^1 in x, u, and t, subject to $\Psi(t_f) = 0$. We consider the task

$$\min_{v,u} J \tag{4}$$

The necessary conditions we obtained in [10] are as follow:

Theorem Consider the system (1) with the performance criterion (3) and assumed that the optimal control actions exists. Let v^* and u^* denote the optimal values of v and u, then v^* and u^* satisfy the following equations:

$$\begin{aligned}
H(x,u,\lambda,t) &= L(x,u,t) + \lambda^T f(x,u), \\
&\ \min_{|u|\leq 1} H(x,u,\lambda,t), \\
\dot{\lambda} &= -H_x^T; \\
\lambda(t_{i+}) &= (I - \frac{dh(x(t_{i+}))}{dx(t_{i+})} c_i)^T \lambda(t_i); \\
\lambda(t_f) &= \Phi_{xf}^T; \\
H(t_i) - H(t_{i+}) &= 0; \\
\text{or } t_1 &= t_0; \\
\text{or } t_{i+1} &= t_i + T; \\
H(t_f) &= 0; \\
t_{k+1} &= t_f \text{ (here if } t_f \text{ is fixed)}
\end{aligned}$$

$$\begin{aligned} c_i &= -\gamma sign(\lambda(t_i)^T h(x(t_{i+})); \\ \text{or}\lambda(t_i)^T h(x(t_{i+})) &= 0; \\ \dot{x} &= f(x,u); \\ x(t_{i+}) &= x(t_i) + h(x(t_{i+}))c_i; \\ i &= 1,2,3,..k+1; \\ \Phi(x(t_f)) &= \phi(x(t_f)) + \nu\Psi(x(t_f)); \\ \Psi(x(t_f)) &= 0; \end{aligned} \tag{5}$$

By using these necessary conditions, the main results on how to construct the switch curves for basic second order systems will be given in next section.

3 General Form Examples and Associated Main Results

The results above can be used to construct the switch curves for time-optimal mixed control action. There are three switch curves to be constructed. They are the optimal impulse firing line (I), its magnitude sign line (II) and bang-bang control switch line (III). First consider just the pure impulse case (set $u \equiv 0$) for some basic nonlinear second order control process.

(1) Example 1 To set notation and establish the setting for feedback impulses, consider double integrator system as follows.

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= \frac{v}{2}, \\ x_1(0) &= x_{10},\ x_2(0) = x_{20}, \\ v &= \sum_i c_i \delta(t - t_i), \\ |c_i| &\leq 1,\ t_{i+1} - t_i > \frac{1}{8}. \end{aligned} \tag{6}$$

The time-optimal task is to select impulses v to take the system's state from original state (x_{10}, x_{20}) to origin as quickly as possible.

By using theorem in section 2, we have adjoint system as follows:

$$\begin{aligned} \dot{\lambda}_1 &= 0, \\ \dot{\lambda}_2 &= -\lambda_1 \end{aligned} \tag{7}$$

Since the optimal impulses will fire when $H(t_i) - H(t_{i+}) = 0$, their magnitude signs will change when $\lambda_2(t) = 0$, set $\lambda_1^2(0) + \lambda_2^2(0) = 1$, and solve the state equations and adjoint system backwards in time from origin. We get switch curves as shown in figure 1. Notice if the impulse fire when above the switch curve (II), then $c_i = -1$, otherwise $c_i = 1$. Using these switch curves (I)(II) and considering that $t_{i+1} - t_i > \frac{1}{8}$, we get the system's trajectories driven by optimal feedback impulses as shown in figure 2 from the initial state $(3.465, 4.500)$. For a general initial state x_{10}, x_{20} and state of recharging for impulse firing we can synthesis a feedback controller with impulse spacing $t_{i+1} - t_i > \frac{1}{8}$ by measuring $x_1(t), x_2(t)$ and the state of recharging. Also the switching may lead to to a sliding mode approach to a neighborhood of the origin.

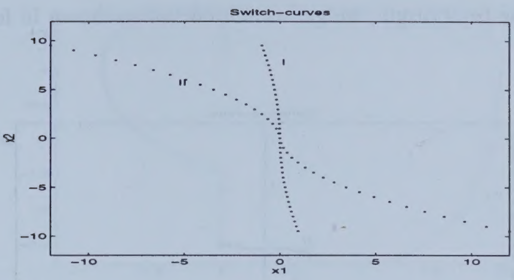


Figure 1: $|c_i| = 1$, synthesis of optimal impulse actions.

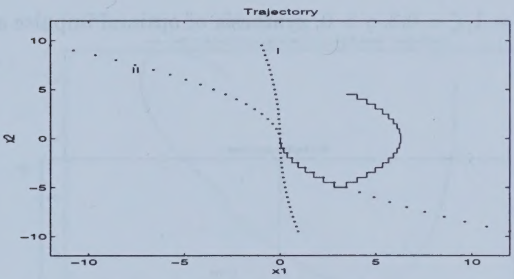


Figure 2: $|c_i| \leq 1$, system's trajectory driven by optimal feedback impulses for $t_{i+1} - t_i > \frac{1}{8}$ at$(3.465, 4.500)$.

(2) Example 2 The nonlinear damped harmonic oscillator has state model

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 - \epsilon x_1^3 - 2\zeta x_2 + v, \\ x_1(0) &= x_{10},\ x_2(0) = x_{20}, \end{aligned} \tag{8}$$

where $\epsilon > 0$ denotes coefficient of nonlinear term and $0 \leq \zeta < 1$ is damping coefficient and v is the impulsive control action.

The goal for choice of optimal impulse actions $v(t) = \sum c_i \delta(t - t_i)$, $|c_i| \leq \gamma$ can be associated with index

$$J = \min_{v} \int_0^{t_f} dt = t_f(v). \tag{9}$$

The amplitude limit γ will be specified below in terms of a total energy rate per cycle. By using the theorem in section 2, we have the adjoint system as follows:

$$\begin{aligned} \dot{\lambda}_1 &= (1 + 3\epsilon x_1^2)\lambda_2, \\ \dot{\lambda}_2 &= -\lambda_1 = 2\zeta\lambda_2 \end{aligned} \tag{10}$$

The optimal impulses will fire when $H(t_i) - H(t_{i+}) = 0$, their magnitude signs will change when $\lambda_2(t) = 0$, set $\lambda_1^2(0) + \lambda_2^2(0) = 1$, and solve the state equations (8) and adjoint system

(10)backwards in time from origin, we get switch curves as shown in following figures 3 and 4.

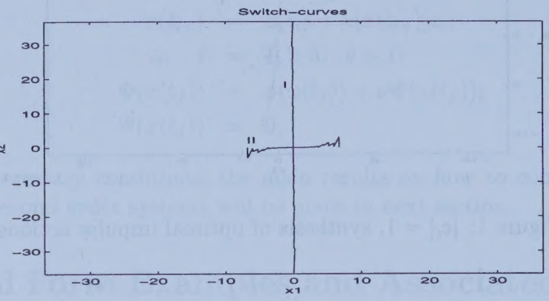


Figure 3: $\epsilon = 1, \zeta = 0.3, \gamma > 0$, synthesis of optimal impulse actions.

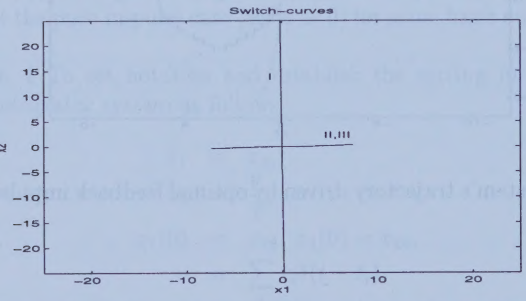


Figure 4: $\epsilon = 0.1, \zeta = 0.8, \gamma > 0$, synthesis of optimal impulse actions.

Using these switch curves in figures 3 and 4 while keeping the total sum of absolute values of c_i to 2 in one period of time, we obtain the system's trajectories driven by optimal feedback impulses as shown in figures 5 and 6. Notice if the impulse fire when above the switch curve (II), then $c_i = -1$, otherwise $c_i = 1$.

Remarks: (1) For the constraint of impulse time spacing t_i being $T > 0$, the optimal impulse firing lines are still the same as above. A trajectory of the system from one point of the switch curve to next meeting of the switch curve can happen without an impulse firing if the recharging condition $t_{i+1} - t_i \geq T$ is enforced. So again the state of the recharging must be used in the decision for impulse firing.

(2) The constraint of the impulse time spacing t_i may cause the optimal impulse actions happening at exactly $t_{i+1} - t_i = T$. In general, except for the final impulse in the sequence, both control variables use the maximum energy available to reach the target set in minimum time.

In the rest of this section, consider mixed control action (which means $|u| \leq \alpha$) and give the switch curves for both u and v. An example is as follows.

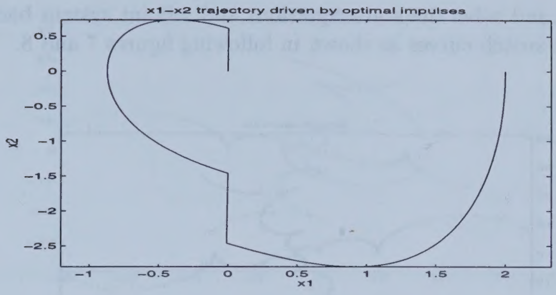


Figure 5: $\epsilon = 1, \zeta = 0.3, \gamma \leq 1$, system's trajectory driven by optimal feedback impulses at$(2, 0)$.

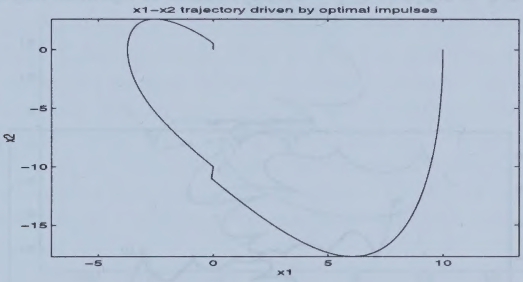


Figure 6: $\epsilon = 0.1, \zeta = 0.8, \gamma \leq 1$, system's trajectory driven by optimal feedback impulses at$(10, 0)$.

(3) Example 3 is the nonlinear damped harmonic oscillator of example two with the added control u:

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 - \epsilon x_1^3 - 2\zeta x_2 + u + v, \\ x_1(0) &= x_{10},\ x_2(0) = x_{20}, \\ |u| &\leq \alpha. \end{aligned} \tag{11}$$

The goal for choice of optimal impulse actions $v(t) = \sum c_i \delta(t - t_i)$, $|c_i| \leq \gamma$ can be associated with index

$$J = \min_{u,v} \int_0^{t_f} dt = t_f(u, v). \tag{12}$$

By using theorem above, we have adjoint system as follows:

$$\begin{aligned} \dot{\lambda}_1 &= (1 + 3\epsilon x_1^2)\lambda_2, \\ \dot{\lambda}_2 &= -\lambda_1 = 2\zeta\lambda_2 \end{aligned} \tag{13}$$

By the necessary conditions $u(t) = \alpha sign(\lambda_2(t))$ and the optimal impulses will fire when $H(t_i) - H(t_{i+}) = 0$ (here the switch curve (II) is the same as switch curve (III)), set

$\lambda_1^2(0) + \lambda_2^2(0) = 1$, and solve the state equations and adjoint system backwards in time from origin, we get switch curves as shown in following figures 7 and 8.

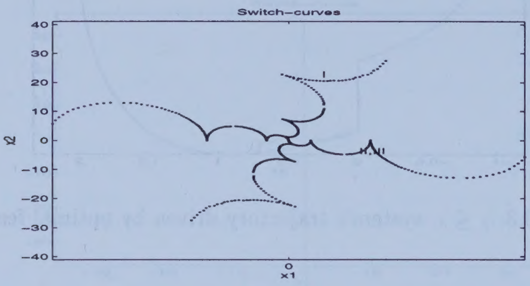


Figure 7: $\alpha = 1, \zeta = 0.3, \epsilon = 0, \gamma = 0.1$, synthesis of optimal impulse actions.

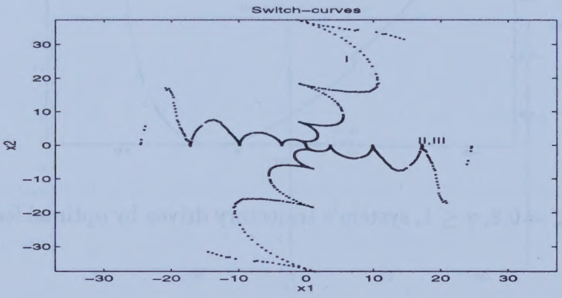


Figure 8: $\alpha = 1, \zeta = 0.3, \epsilon = 0.01, \gamma = 0.1$, synthesis of optimal impulse actions.

The optimal impulse firing line changes from straight line as in figures 3 and 4 based on pure impulse actions to the shapes shown in figures 7 and 8 and keeping the total sum of absolute values of c_i to 0.2 during one period of time, we get the system's trajectories driven by optimal feedback impulses as shown in figures 9 and 10. Also notice if the impulse fire when above the switch curve (II), then $c_i = -1$, otherwise $c_i = 1$.

4 Conclusions

This paper presents the solutions to the optimization tasks of impulse control combined with usual bounded control. The results may be considered in the setting of a generalized Maximal principle.

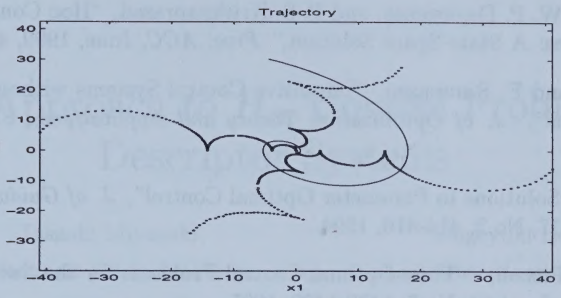


Figure 9: $\alpha = 1, \zeta = 0.3, \epsilon = 0, \gamma = 0.1$, system's trajectory driven by optimal feedback impulses at$(-4.4189, 30.1782)$.

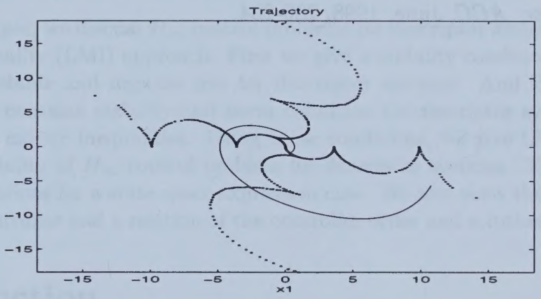


Figure 10: $\alpha = 1, \zeta = 0.3, \epsilon = 0.01, \gamma = 0.1$, system's trajectory driven by optimal feedback impulses at$(10.7629 - 5.6602,)$.

References

[1] W.W.Schmaedeke, " Optimal Control Theory for Nonlinear Vector Differential Equations Containing Measures," *J. of Ind. Appl. Math., Ser. A, Control 3, No.2* 231-280,1965

[2] E.B.Lee and L. Markus, Foundations of Optimal Control Theory, Wiley, N. Y., 1967.

[3] An-Hwa Chang, "Optimal Control with Mixed Control Actions," *Ph.D Thesis*, U. of M. Feb. 1995.

[4] E.B.Lee And J.C. Luo "Two methods for constructing aggrandized disturbances using delay tactics for stabilized control systems," *Proc. ACC*, June,1997, 2189-2192.

[5] Boris M. Miller, "The Generalized solutions of Nonlinear Optimization Problems with Impulse," *J. of SIAM Control and Optimization*, Vol.34 No.4 1420-1440, July,1996.

[6] Q. F. Wei, W. P. Dayawansa, and P. S. Krishnaprasad, "H∞ Control for Impulsive Disturbances: A State-Space Solution," *Proc. ACC*, June, 1995, 4379-4383.

[7] A. Bressa and F. Rampazzo, "Impulsive Control Systems without Commutativity Assumptions", *J. of Optimization Theory and Applications,*, 81, No.3, 435-457, 1994.

[8] Y. Zhao, " Solutions to Parameter Optimal Control", *J. of Guidance, Control, and Dynamics*, 17, No.2, 414-416, 1994.

[9] Benedetto Piccoli, " Time-Optimal Control Problems for the Swing and Ski", *Int. J. of Control*, vol.62, No.6, 1409-1429, 1995.

[10] E.B.Lee And J.C. Luo "The necessary Conditions of Nonlinear Optimization with Impulse control (Or Impulse Disturbance)," to be published.

[11] J.C. Luo and E.B.Lee " Time-Optimal Cotrol of the Swing using Impulse Cotrol Actions", *Proc. ACC*, June, 1998, 200-204.

A LMI Approach to H_∞ Control Problems for Descriptor Systems

Takashi Miyazaki
Dept. of Electronic-Mechanical Engineering
Nagoya University, Japan
miyazaki@nuem.nagoya-u.ac.jp

Shigeyuki Hosoe
hosoe@nuem.nagoya-u.ac.jp

Abstract

In this paper, we discuss H_∞ control problems for descriptor systems via Linear Matrix Inequality (LMI) approach. First we give a stability condition that means exponential stable and impulse free for descriptor systems. And Bounded Real Lemma that contains stability and norm condition for descriptor systems can be expressed as matrix inequalities. Using these conditions, we give LMI conditions for the solvability of H_∞ control problem for descriptor systems. The results are natural extensions for a state space expression case. We also show the construction of proper controller and a relation of the controller order and solutions of the LMI.

1 Introduction

Various approaches have already been developed for solving H_∞ control problems based on a state space representation. They can be classified into a Riccati equation approach [2, 3] and a matrix inequalities [5, 4] approach. The former approach had an advantage in numerical calculations, since Riccati equations can be solved by reducing them into an eigenvalue problem. But to follow the approach we need some restrictive assumptions about zeros of a generalized plant. On the other hand, the matrix inequalities approach did not have to satisfy such assumptions, but had some difficulty for efficient numerical solution method. Recently, however solvability conditions of H_∞ control problems using Linear Matrix Inequalities (LMI, [7, 8]) has been developed with efficient numerical algorithms [9, 10].

There also are many researches of control problems based on descriptor form systems [1], which is a natural representation of physical systems. The descriptor system has the advantage that in the representation physical parameters can be included without large modifications and even the systems that contain static constraints and impulsive elements can be included.

In this paper, we discuss H_∞ control problems for descriptor systems via LMI approach. In [6], Masubuchi et al. showed solvability conditions of this problem using Riccati inequalities and reduced them to LMI conditions. They also got a controller that has same size as the generalized plant. We will show a solvability condition directly based on LMI approach and get a similar result concerning the order of controllers which is known for the state space representation.

Notations:

$$\left[\begin{array}{c|c} -sE+A & B \\ \hline C & D \end{array}\right] := C\,(sE-A)^{-1}\,B + D$$

2 Stability

The stability of descriptor systems is defined as follows.

Definition 1 *For $E, A \in \mathbf{R}^{n\times n}$, we call that a pair (E, A) is stable if the following conditions are satisfied:*

1. *Regularity*
$$\det\left(-sE+A\right) \not\equiv 0$$
2. *Exponential stability*
$$\det\left(-sE+A\right) \neq 0,\ \forall \Re\left(s\right) \geq 0$$
3. *Impulse free*
$$ImE + A\ker E = \mathbf{R}^n.$$

This stability condition can be described by the following LMI condition.

Lemma 1 ([6]) *(E, A) is stable if and only if there exists a solution $X \in \mathbf{R}^{n\times n}$ of equations*

$$E^TX = X^TE \geq 0 \tag{1}$$

$$A^TX + X^TA < 0 \tag{2}$$

where $A,\ E \in \mathbf{R}^{n\times n}$.

Proof

We provide a direct proof that is different from Ref. [6].

The conditions (1) and (2) are preserved under the transformations

$$-s\bar{E} + \bar{A} = T_l\left(-sE+A\right)T_r,\quad \bar{X} = T_l^{-T}XT_r$$

by using nonsingular matrices T_l and T_r. Therefore we can assume that the regular pencil $-sE+A$ is Weierstrass normal form

$$-sE+A = \begin{bmatrix} -sI+A_1 & 0 \\ 0 & -sN+I \end{bmatrix}$$

where N is a nilpotent matrix. According to this, denote X as

$$X = \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix}.$$

Then (1) and (2) become as

$$\begin{bmatrix} X_1 & X_2 \\ N^TX_3 & N^TX_4 \end{bmatrix} = \begin{bmatrix} X_1^T & X_3^TN \\ X_2^T & X_4^TN \end{bmatrix} \geq 0 \tag{3}$$

$$\left[\begin{array}{cc} A_1^T X_1 + X_1^T A_1 & X_3^T + A_1^T X_2 \\ X_3 + X_2^T A_1 & X_4^T + X_4 \end{array}\right] < 0. \tag{4}$$

[Sufficiency] Assume that (1) and (2) are satisfied. From (2) A is nonsingular, $(E,\ A)$ is regular and X_1 is positive definite matrix. Therefore the left upper blocks of (3) and (4) show that A_1 is Hurbitz matrix and $(E,\ A)$ is exponential stable. Finally we show that $(E,\ A)$ is impulse free i.e. $N = 0$. If we assume that $N \neq 0$, there exists an integer $\alpha \geq 2$ such that $N^{\alpha-1} \neq 0$ and $N^{\alpha} = 0$. Since $N^T X_4 = X_4^T N$ from (3), it follows that

$$\begin{aligned} \left(N^{\alpha-1}\right)^T \left(X_4 + X_4^T\right) N^{\alpha-1} &= \left(N^{\alpha-2}\right)^T X_4^T N^{\alpha} + \left(N^{\alpha}\right)^T X_4 N^{\alpha-2} \\ &= 0. \end{aligned}$$

Because $X_4 + X_4^T < 0$, this means that $N^{\alpha-1} = 0$ and contradicts the assumption. Therefore we can conclude that $N = 0$.

[Necessity] It is obvious. We omit the proof.

Next we discuss a bounded real condition for a descriptor system

$$\begin{aligned} E\dot{x} &= Ax + Bu \\ y &= Cx. \end{aligned}$$

Notice that we assume without loss of generality that there is no (apparent) direct term from u to y i.e. $D = 0$. This assumption is useful when describing the bounded-real condition based on LMI.

Lemma 2 *$(E,\ A)$ is stable and $\left\|C\left(sE - A\right)^{-1} B\right\|_{\infty} < \gamma$ if and only if there exits $X \in \mathbf{R}^{n \times n}$ such that*

$$E^T X = X^T E \geq 0 \tag{5}$$

$$\left[\begin{array}{ccc} A^T X + X^T A & X^T B & C^T \\ B^T X & -\gamma I & 0 \\ C & 0 & -\gamma I \end{array}\right] < 0 \tag{6}$$

where $A,\ E \in \mathbf{R}^{n \times n}$.

Proof

The conditions (5) and (6) have a different form from the ones of Ref.[6], but they are equivalent. Now we provide an alternative proof by using a LMI expression.

Notice first that Lemma 1 with (5) and a upper-left part of (6) implies that $(E,\ A)$ is stable.

In the following, as in the proof of Lemma 1, let us assume without loss of the generality that the system pencil is already in the Weierstrass form

$$\left[\begin{array}{cc} -sE + A & B \\ C & 0 \end{array}\right] = \left[\begin{array}{cc|c} -sI + A_1 & 0 & B_1 \\ 0 & I & B_2 \\ \hline C_1 & C_2 & 0 \end{array}\right].$$

We write

$$X = \left[\begin{array}{cc} X_1 & X_2 \\ X_3 & X_4 \end{array}\right]$$

according to this partitioning. Then a transfer function matrix becomes

$$C\left(sE - A\right)^{-1} B = C_1\left(sI - A_1\right)^{-1} B_1 - C_2 B_2, \tag{7}$$

and (5) is equivalently rewritten as

$$X_1 = X_1^T > 0, \quad X_2 = 0.$$

Multiplying a matrix

$$\begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & -B_2 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}$$

from the right side of (6) and its transpose matrix from the left side, (6) becomes as

$$\begin{bmatrix} \Psi_1 & X_3^T & \Psi_3 \\ X_3 & \Psi_2 & \Psi_4 \\ \Psi_3^T & \Psi_4^T & \Psi_5 \end{bmatrix} < 0 \tag{8}$$

where

$$\Psi_1 = A_1^T X_1 + X_1 A_1, \quad \Psi_2 = X_4 + X_4^T, \quad \Psi_3 = \begin{bmatrix} X_1 B_1 & C_1^T \end{bmatrix}, \quad \Psi_4 = \begin{bmatrix} -X_4 B_2 & C_2^T \end{bmatrix},$$

$$\Psi_5 = \begin{bmatrix} -\gamma I & -B_2^T C_2^T \\ -C_2 B_2 & -\gamma I \end{bmatrix}.$$

[Necessity] If the system is stable and the norm condition is satisfied, then from the results of state space representation and the transfer function matrix (7) there exists $X_1 > 0$ such that

$$\begin{bmatrix} \Psi_1 & \Psi_3 \\ \Psi_3^T & \Psi_5 \end{bmatrix} < 0. \tag{9}$$

Since it is also satisfied that $\|C_2 B_2\|_\infty < \gamma$, $\left\| C_2 (sI + I)^{-1} B_2 - C_2 B_2 \right\|_\infty < \gamma$ and there exists $\bar{X}_4 = \bar{X}_4^T > 0$ such that

$$\begin{bmatrix} -\bar{X}_4 - \bar{X}_4 & \bar{X}_4 B_2 & C_2^T \\ B_2^T \bar{X}_4 & -\gamma I & -B_2^T C_2^T \\ C_2 & -C_2 B_2 & -\gamma I \end{bmatrix} < 0.$$

It follows

$$\begin{bmatrix} \Psi_2 & \Psi_4 \\ \Psi_4^T & \Psi_5 \end{bmatrix} < 0$$

where $X_4 = -\bar{X}_4$. Then, from the above two inequalities

$$\begin{bmatrix} \Psi_1 & \Psi_3 & 0 \\ \Psi_3^T & \Psi_5 & 0 \\ 0 & 0 & \Psi_2 - \Psi_4 \Psi_5^{-1} \Psi_4^T \end{bmatrix} < 0.$$

Then, if we multiply matrix

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & \Psi_5^{-1} \Psi_4^T \\ 0 & 0 & I \end{bmatrix}$$

from the right side of this inequality and its transpose matrix from the left side and letting $X_3 = \Psi_4 \Psi_5^{-1} \Psi_3^T$, we get (8).

[Sufficiency] Assume that (5) and (6) are satisfied. Then (8) holds and therefore (9) follows which is a bounded real condition for a state space representation.

3 H_∞ Control Problem

We consider the following continuous time linear time-invariant generalized plant represented by the descriptor form

$$\begin{aligned} E\dot{x} &= Ax + B_1 w + B_2 u \\ z &= C_1 x \\ y &= C_2 x \end{aligned} \tag{10}$$

where $w \in \mathbf{R}^{m_1}$ is the exogenous input, $u \in \mathbf{R}^{m_2}$ is the control input, $z \in \mathbf{R}^{p_1}$ is the controlled output and $y \in \mathbf{R}^{p_2}$ is the measured output.

Let $T_{zw}(s)$ be a transfer function matrix from w to z. The H_∞ control problem is formulated to find a output feedback control law $u = K(s)y$ such that for a given γ, a norm condition $\|T_{zw}(s)\|_\infty < \gamma$ is satisfied and the closed-loop system is internally stable.

The following theorem is our main results.

Theorem 1 *For the given system (10) and $\gamma > 0$, there exists a controller satisfying the norm condition and the internal stability if and only if there exist $R,\ S \in \mathbf{R}^{n\times n}$ such that*

$$\begin{bmatrix} E & 0 \\ 0 & E^T \end{bmatrix}\begin{bmatrix} R & I \\ I & S \end{bmatrix} = \begin{bmatrix} R^T & I \\ I & S^T \end{bmatrix}\begin{bmatrix} E^T & 0 \\ 0 & E \end{bmatrix} \geq 0 \tag{11}$$

$$\begin{bmatrix} B_2^\perp & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}\begin{bmatrix} AR + R^T A^T & B_1 & R^T C_1^T \\ B_1^T & -\gamma I & 0 \\ C_1 R & 0 & -\gamma I \end{bmatrix}\begin{bmatrix} B_2^{\perp T} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} < 0 \tag{12}$$

$$\begin{bmatrix} C_2^{\perp T} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}\begin{bmatrix} A^T S + S^T A & S^T B_1 & C_1^T \\ B_1^T S & -\gamma I & 0 \\ C_1 & 0 & -\gamma I \end{bmatrix}\begin{bmatrix} C_2^\perp & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} < 0 \tag{13}$$

where $B_2^\perp$ and $C_2^{\perp T}$ are respectively maximum dimension matrices such that

$$B_2^\perp \begin{bmatrix} B_2 & B_2^{\perp T} \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix},\ C_2^{\perp T}\begin{bmatrix} C_2^T & C_2^\perp \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix}.$$

Moreover, there exists a proper controller $K(s)$ of order $\bar{k}$ if and only if there exist R and S that satisfy (11),(12),(13) and

$$\bar{k} \geq \operatorname{rank} E\,(I - RS)\,. \tag{14}$$

Proof Let a controller be represented by

$$K(s) = \left[\begin{array}{c|c} -sE_k + A_k & B_k \\ \hline C_k & D_k \end{array}\right] \tag{15}$$

where $E_k, A_k \in \mathbf{R}^{k\times k}$,$B_k \in \mathbf{R}^{k\times p_2}$, $C_k \in \mathbf{R}^{m_2\times k}$,$D_k \in \mathbf{R}^{m_2\times p_2}$. We define

$$\Theta = \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix},\ \hat{E} = \begin{bmatrix} E & 0 \\ 0 & E_k \end{bmatrix},\ \hat{A} = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}$$

$$\hat{B}_1 = \begin{bmatrix} B_1 \\ 0 \end{bmatrix},\ \hat{B}_2 = \begin{bmatrix} 0 & B_2 \\ I & 0 \end{bmatrix},\ \hat{C}_1^T = \begin{bmatrix} C_1^T \\ 0 \end{bmatrix},\ \hat{C}_2^T = \begin{bmatrix} 0 & C_2^T \\ I & 0 \end{bmatrix}.$$

The transfer function of the closed loop system from w to z is represented as follows:

$$T_{zw} = \left[\begin{array}{c|c} -s\hat{E} + \hat{A} + \hat{B}_2\Theta\hat{C}_2 & \hat{B}_1 \\ \hline \hat{C}_1 & 0 \end{array}\right].$$

From Lemma 2, the solvability of H_∞ control problem is equivalent to the existence of $X \in \mathbf{R}^{(n+k)\times(n+k)}$ such that

$$\hat{E}^T X = X^T \hat{E} \geq 0 \tag{16}$$

$$\Psi(X) + Q^T\Theta^T\hat{P}(X) + \hat{P}^T(X)\Theta Q < 0 \tag{17}$$

where

$$\hat{P}(X) = \begin{bmatrix} \hat{B}_2^T X & 0 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} \hat{C}_2 & 0 & 0 \end{bmatrix}$$

$$\Psi(X) = \begin{bmatrix} \hat{A}^T X + X^T \hat{A} & X^T \hat{B}_1 & \hat{C}_1^T \\ \hat{B}_1^T X & -\gamma I & 0 \\ \hat{C}_1 & 0 & -\gamma I \end{bmatrix}.$$

We define matrices

$$P = \begin{bmatrix} \hat{B}_2^T & 0 & 0 \end{bmatrix}, \quad \Phi(X^{-1}) = \begin{bmatrix} X^{-T}\hat{A}^T + \hat{A}X^{-1} & \hat{B}_1 & X^{-T}\hat{C}_1^T \\ \hat{B}_1^T & -\gamma I & 0 \\ \hat{C}_1 X^{-1} & 0 & -\gamma I \end{bmatrix}$$

and let $P^{\perp T}$ and $Q^{\perp T}$ be respectively maximum dimension matrices such that

$$P^{\perp T}\begin{bmatrix} P^T & P^\perp \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix}, \quad Q^{\perp T}\begin{bmatrix} Q^T & Q^\perp \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix}.$$

Then the solvability of (17) with respect to Θ is equivalent [7] to the conditions

$$P^{\perp T}\Phi(X^{-1})P^\perp < 0 \tag{18}$$

$$Q^{\perp T}\Psi(X)Q^\perp < 0. \tag{19}$$

After all, the solvability of H_∞ control problem is equivalent to the existence of X and E_k that satisfy (16), (18) and (19).

[Necessity] Suppose that there exists such X and E_k. According to the partition of $\hat{E}$, let

$$X = \begin{bmatrix} S & S_2 \\ S_3 & S_4 \end{bmatrix}, \quad X^{-1} = \begin{bmatrix} R & R_2 \\ R_3 & R_4 \end{bmatrix}. \tag{20}$$

We can take

$$P^{\perp T} = \begin{bmatrix} B_2^\perp & 0 & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}.$$

Then from (18), (12) follows. In the same way, (19) with

$$Q^{\perp T} = \begin{bmatrix} C_2^{\perp T} & 0 & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}$$

gives (13).

From (16), we have

$$\begin{bmatrix} E^T S & E^T S_2 \\ E_k^T S_3 & E_k^T S_4 \end{bmatrix} = \begin{bmatrix} S^T E & S_3^T E_k \\ S_2^T E & S_4^T E_k \end{bmatrix} \geq 0 \tag{21}$$

$$\begin{bmatrix} R^T E^T & R_3^T E_k^T \\ R_2^T E^T & R_4^T E_k^T \end{bmatrix} = \begin{bmatrix} ER & ER_2 \\ E_k R_3 & E_k R_4 \end{bmatrix} \geq 0 \tag{22}$$

and the relation of inverse matrices gives

$$\begin{bmatrix} SR + S_2 R_3 & SR_2 + S_2 R_4 \\ S_3 R + S_4 R_3 & S_3 R_2 + S_4 R_4 \end{bmatrix} = I \tag{23}$$

$$\begin{bmatrix} RS + R_2 S_3 & RS_2 + R_2 S_4 \\ R_3 S + R_4 S_3 & R_3 S_2 + R_4 R_4 \end{bmatrix} = I. \tag{24}$$

Using these equations, we get

$$\begin{bmatrix} I & 0 \\ S^T & S_3^T \end{bmatrix} \begin{bmatrix} ER & ER_2 \\ E_k R_3 & E_k R_4 \end{bmatrix} \begin{bmatrix} I & S \\ 0 & S_3 \end{bmatrix} = \begin{bmatrix} ER & E \\ E^T & E^T S \end{bmatrix}. \tag{25}$$

This proves (11).

If there exists a proper controller of order $\bar{k}$, we can suppose $E_k = I_k$ and $k = \bar{k}$. Because $S_3 \in \mathbf{R}^{\bar{k} \times n}$ if $\bar{k} \leq n$, we have $\bar{k} \geq \operatorname{rank} S_3$. In this case, since $I - RS = R_2 S_3$ from (24), (14) follows. If $\bar{k} > n$, (14) is obvious.

[Sufficiency] Assume that there exist R and S that satisfy (11), (12) and (13). If we partition X and X^{-1} as (20), (12) and (13) lead to (18) and (19) respectively.

To complete the proof, we shall construct R_2, R_3, R_4, S_2, S_3, S_4 and E_k that satisfy (21)-(24).

At first we consider the case $k = \operatorname{rank}(I - RS)$. Then there exist full column rank matrices $R_2, R_3^T, S_2, S_3^T \in \mathbf{R}^{n \times k}$ such that

$$S_2 R_3 = I - SR, \quad R_2 S_3 = I - RS. \tag{26}$$

Note

$$E(I - RS) = ER_2 S_3 = \left(I - R^T S^T\right) E = R_3^T S_2^T E.$$

Then we have

$$\ker S_3 \subseteq \ker ER_2 S_3 = \ker S_2^T E$$

Thus there exists $E_k \in \mathbf{R}^{k \times k}$ such that

$$E_k^T S_3 = S_2^T E. \tag{27}$$

Because R_3 is of full row rank, it follows that $\operatorname{rank} S_2^T E = \operatorname{rank} E(I - RS)$. Moreover since S_3 is of full row rank, it follows that

$$\operatorname{rank} E_k = \operatorname{rank} E(I - RS). \tag{28}$$

Then since $\left(ER_2 - R_3^T E_k^T\right) S_3 = 0$, $ER_2 = R_3^T E_k^T$. From $\ker R_2^T \subseteq \ker R^T S_3^T R_2^T = \ker R^T (RS - I)^T$, $\ker R_3 \subseteq \ker RS_2 R_3 = \ker (RS - I) R$, there exists $S_4 \in \mathbf{R}^{k \times k}$ such that

$$R_2 S_4 R_3 = (RS - I) R. \tag{29}$$

Similarly there exists $R_4 \in \mathbf{R}^{k \times k}$ such that $S_2 R_4 S_3 = (SR - I)\, S$. Because

$$\begin{aligned} R_3^T E_k^T S_4 R_3 &= E\,(RS - I)\, R \\ &= R^T \left(S^T R^T - I \right) E^T \\ &= R_3^T S_4^T E_k R_3 \end{aligned}$$

holds and R_3 is a full row rank matrix, we have $E_k^T S_4 = S_4^T E_k$. In the same way, we can show that $R_4^T E_k^T = E_k R_4$. Since

$$R_2\,(S_3 R + S_4 R_3) = 0, \;\; (S R_2 + S_2 R_4)\, S_3 = 0,$$

it follows that $S_3 R + S_4 R_3 = 0$, $S R_2 + S_2 R_4 = 0$. Similarly we can get $R_3 S + R_4 S_3 = 0$, $R S_2 + R_2 S_4 = 0$. Because

$$\begin{aligned} (S_3 R_2 + S_4 R_4)\, S_3 &= S_3 R_2 S_3 - S_4 R_3 S \\ &= S_3\,(I - RS) + S_3 RS \\ &= S_3, \end{aligned}$$

it follows that $S_3 R_2 + S_4 R_4 = I$. In the same way, we can show that $R_3 S_2 + R_4 S_4 = I$. From these equations, (25) follows. Because the positive semi-definite condition of (11) is satisfied and S_3 is full row rank, the positive semi-definite condition of (22) holds. Similarly we can show the positive semi-definite condition of (21). Hence we showed (21)-(24).

In the case $k > \operatorname{rank}(I - RS)$, it is sufficient to construct X and E_k as follows. Let $\tilde{k} = \operatorname{rank}(I - RS)$. Then following the above discussion we can get $\tilde{R}_2$, $\tilde{R}_3$, $\tilde{R}_4$, $\tilde{S}_2$, $\tilde{S}_3$, $\tilde{S}_4$ and $\tilde{E}_k$ and construct the following matrices

$$X = \left[\begin{array}{c|cc} S & \tilde{S}_2 & 0 \\ \hline \tilde{S}_3 & \tilde{S}_4 & 0 \\ 0 & 0 & I_{k-\tilde{k}} \end{array} \right], \quad X^{-1} = \left[\begin{array}{c|cc} R & \tilde{R}_2 & 0 \\ \hline \tilde{R}_3 & \tilde{R}_4 & 0 \\ 0 & 0 & I_{k-\tilde{k}} \end{array} \right], \quad E_k = \begin{bmatrix} \tilde{E}_k & 0 \\ 0 & I_{k-\tilde{k}} \end{bmatrix}.$$

By using these X and E_k, we can get solution Θ to (17). Then controller $K(s)$ is determined by E_k and Θ. But with this results we can not guarantee the properness of the controller. To consider the existence of a proper controller, let $\bar{k} = \operatorname{rank} E\,(I - RS)$. From (28) there exist nonsingular matrices T_L and T_R such that

$$T_L\,(-sE_k + A_k)\, T_R = \begin{bmatrix} -sI_{\bar{k}} + A_{k_1} & A_{k_2} \\ A_{k_3} & A_{k_4} \end{bmatrix}.$$

Also for any real number $\epsilon > 0$ there exists a matrix $\bar{A}_{k_4}$ such that $A_{k_4} + \epsilon \bar{A}_{k_4}$ is nonsingular. Define the following matrices

$$\bar{A}_k = T_L^{-1} \begin{bmatrix} 0 & 0 \\ 0 & \bar{A}_{k_4} \end{bmatrix} T_R^{-1}, \;\; \bar{\Theta} = \begin{bmatrix} \bar{A}_k & 0 \\ 0 & 0 \end{bmatrix}.$$

Because of (17) there exists a ϵ such that

$$\Psi + Q^T \Theta^T \hat{P} + \hat{P}^T \Theta Q + \epsilon \left(Q^T \bar{\Theta}^T \hat{P} + \hat{P}^T \bar{\Theta} Q \right) < 0.$$

From the construction the controller

$$\bar{K}(s) = \left[\begin{array}{c|c} -sE_k + A_k + \epsilon \bar{A}_k & B_k \\ \hline C_k & D_k \end{array} \right] \tag{30}$$

is a solution of H_∞ control problem and this is proper controller. The order of the transfer function is less than $\bar{k}$.

These conditions is a natural extension to the results of the state space expression [7, 8]. We could provide the LMI conditions directly, differently from Ref. [6]. We could also show a similar result concerning the order of controller as in the state space expression case.

4 Consideration

Sometimes, equality (11) is not tractable in a numerical calculation. To avoid this we can use the following equivalent transformation. First using a singular-value decomposition, put E as

$$\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} E \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$$

where $\Sigma > 0$. Let $\hat{R}_a$, $\hat{R}_b$, $\hat{S}_a$ and $\hat{S}_b$ be new variables. We fix the variables as

$$R = \hat{R}_a E^T + V_2 \hat{R}_b, \quad \hat{R}_a = \hat{R}_a^T \tag{31}$$

$$S = \hat{S}_a E + U_2^T \hat{S}_b, \quad \hat{S}_a = \hat{S}_a^T \tag{32}$$

and replace (11) by

$$\begin{bmatrix} E\hat{R}_a E^T & E \\ E^T & E^T \hat{S}_a E \end{bmatrix} \geq 0. \tag{33}$$

Notice that these reductions do not lose any generality, It can be confirmed as follows.

Assume that there exist R and S that satisfy (11). Define $U^T = \begin{bmatrix} U_1^T & U_2^T \end{bmatrix}$ and $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$, then it follows that

$$V^{-1} R U^T = \begin{bmatrix} \hat{R}_1 & 0 \\ \hat{R}_3 & \hat{R}_4 \end{bmatrix}, \quad U^{-T} S V = \begin{bmatrix} \hat{S}_1 & 0 \\ \hat{S}_3 & \hat{S}_4 \end{bmatrix}$$

$$\hat{R}_1 = \hat{R}_1^T, \quad \hat{S}_1 = \hat{S}_1^T, \quad \begin{bmatrix} \hat{R}_1 & I \\ I & \hat{S}_1 \end{bmatrix} \geq 0.$$

Therefore $\hat{R}_a = V_1 \hat{R}_1 V_1^T$, $\hat{R}_b = \begin{bmatrix} \hat{R}_3 & \hat{R}_4 \end{bmatrix} U^{-T}$ satisfy (31). Matrices $\hat{S}_a = U_1^T \hat{S}_1 U_1$, $\hat{S}_b = \begin{bmatrix} \hat{S}_3 & \hat{S}_4 \end{bmatrix} V^{-1}$ also satisfy (32) and (33) holds.

Finally, let us give some comments on the implication of the assumption in the theorem. Notice first that the conditions of the theorem require that $(E,\ A,\ B_2)$ is impulse controllable i.e.

$$\operatorname{Im} E + A \ker E + \operatorname{Im} B_2 = \mathbf{R}^n.$$

If $(E,\ A,\ B_2)$ is not impulse controllable, then there exists $x \in \mathbf{R}^n (x \neq 0)$ such that

$$x^T \begin{bmatrix} E & B_2 \end{bmatrix} = 0$$

$$x^T A y = 0, \quad \forall y \in \ker E.$$

On the other hand, $ER = R^T E^T$ shows that $Rx \in \ker E$. Therefore

$$x^T \left(AR + R^T A^T \right) x = 0$$

holds. This contradicts (12).

Dually from the assumptions of the theorem we can show that (E, A, C_2) must be impulse observable. Notice that these facts is giving some restrictions on a class of generalized plants and the theorem may be too restrictive for some applicational problems. To remove the restrictions, we have to change the stability requirement a little bit. This problem is being considered now.

5 Conclusion

In this paper, we gave LMI conditions for the solvability of H_∞ control problem for descriptor systems. We also showed the construction of proper controller and a relation of the controller order and solutions of the LMI.

References

[1] F.L. Lewis, "A Survey of Linear Singular Systems," Circuits Syst. Signal Process., 5-1, pp.3-36, 1986.

[2] J. Doyle, K. Glover, P.P. Khargoneker and B.A. Francis, "State Space Solutions to Standard H_2 and H_∞ Control Problems," IEEE Trans., **AC-34**-8, pp.831-847, 1989.

[3] M. Green, K. Glover, D.J.N. Limebeer and J. Doyle, "A J-Spectral Factorization Approach to H_∞ Control," SIAM J. Control and Optimization., **28**-6, pp.1350-1371, 1990.

[4] A.A. Stoorvogel, "The H_∞ Control Problem," Prentice Hall, 1992.

[5] M. Sampei, T. Mita and M. Nakamichi, "An Algebraic Approach to H_∞ output Feedback Control Problems," System and Control Letters, **14**, pp.13-24, 1990.

[6] I. Masubuchi, Y. Kamitane, A.Ohara and N.Suda, "H_∞ Control for Descriptor Systems: A Matrix Inequalities Approach, " Automatica, **33**-4, pp.669-673, 1997.

[7] P. Gahinet and P. Apkarian, "A Linear Matrix Inequality Approach to H_∞ Control," Int. J. Robust and Nonlinear Control, **4**, pp.421-448, 1994.

[8] T. Iwasaki and R. E. Skelton, "All Controllers for the general H_∞ Control Problem: LMI Existence Condition and State Space Formulas," Automatica, **30**, pp.1307-1317, 1994.

[9] S. Boyd and L. El Ghaoui, "Method of Center for Minimizing Generalized Eigenvalues," Linear Algebra and its Applications, **188-189**, pp.63-111, 1993.

[10] Y. Nesterov and A. Nemirovskii, "Interior-point Polynomial Algorithms in Convex Programming," SIAM, Philadelphia, 1994.

Stratified Morse Theory for Stability Margin Computation

Nainn-Ping Ke
Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, CA 90089-2563
email: nke@eudoxus.usc.edu

Abstract

The objective of this paper is to show by using stratified Morse theory the robust stability margin problem can be reformulated as to solve systems of equations. The solutions are the critical points of manifolds. Once all critical points are found, the robust margin can be easily found. Hence, how to solve polynomial systems becomes the central issue for stability margin computation.

1 Introduction

This paper discusses a stability robustness problem by using topological approach. Stability robustness, which every feedback control system should have, in order to remain stable for all conditions, is the most important issue of a feedback controller design. In this paper, we consider the robustness margin problem of plants with real/complex parametric uncertainty. This problem is a standard problem in stability robustness analysis [1, 3, 4, 10]. It is well-known that the real robustness margin can be discontinuous in the problem data while the complex robustness margin is always continuous [3, 21]. In addition, the stability robustness problem with real uncertainty is NP-hard [34]. Hence, many researchers have focused on the robust stability problem with real parametric variations [1, 3, 4, 10, 20]. However, the recently published results show that complex stability robustness problem is also NP-hard [36]. Thus, pursuing exact calculation of a stability robustness margin is a tough job for control engineers. In addition, grid method has been criticized for its troublesome complexity. It is the motivation of this paper to search a new method to calculate the complex robustness margin without using grid method.

The robustness margin problem of plants with parametric uncertainty can be formulated in terms of a canonical diagonal perturbation [10]. Safonov and Athans introduced the multivariable stability margin, k_M, as a measure of stability robustness [35]. An algorithm for exact calculation of the multiloop stability margin with real uncertainties has been proposed by de Gaston and Safonov [10]. Their approach, which is a branch and bound method (i.e. a type of iterative algorithm) requires one to subdivide the uncertainty space in order to have exact k_M in the worst case. However, it provides a good way to apply the Mapping Theorem and the Zero Exclusion Theorem to the stability robustness problem [3, 10]. This approach has recently been modified, extended and labeled by Barmish as the "value set" approach [3]. A number of researcher such as Ackermann, Bhattacharyya, etc. also contributed to the development of the "value set" approach [1, 3, 4].

Whitney and Morse are pioneers of singularity theory, a term which includes catastrophes and bifurcations [5, 32]. Their works were continued and widen by Thom, using transversality as a systematic tool for studying singularities. Then, singularity theory has been developed by Mather, Zeeman, Smale, Arnold and many other researchers for decades [2, 19]. Goresky and MacPheron has established the **Statified Morse theory** [15]. The boundary of the template is contained in the union of critical value curves [20]. In addition, the boundary of the template is first to intercept the origin when k is increased. Hence, the boundary of the template is crucial in the analysis of the robust stability problem [20]. Therefore, the critical value curves play an important role in solving the robust stability problem. Kraus, Anderson and Mansour were first to report the use of singularities for stability analysis [29]. Moreover, the critical point related issues for the robust stability problem have been investigated by a number of another researchers [1, 19, 29]. It has been shown that singularity theory is useful to understand the discontinuity phenomenon [19, 22]. Stratified Morse theory has been applied to solve the stability robustness margin problem by Jonckheere and Ke [20]. Moreover, a unifying mathematical framework, which brings together algebraic topology/differential topology and robust stability, has been developed by Jonckheere in [19]. Symbolic computation has been used to compute real stability margin by Ke in [24]. Complex k_M problem can be reformulated as to solve a set of polynomial equations for determining whether the boundary of the uncertainty is intercept the origin or not by Ke [28]. This paper will extend these results to compute mixed k_M which uncertainty space is defined on a stratified space. The solutions are the critical points of a manifold. Once the critical points are found, the stability margin can be easily found.

2 Problem Formulation

We need to formulate the robust margin problem within a framework amenable to mathematical analysis.

Let $f : \mathbb{D} \times \Omega \ni (d, j\omega) \longmapsto det(I + L(d, j\omega)) \in \mathbb{C}$ be the Nyquist map [19, 27, 26]. The domain $\mathbb{D} \times \Omega$ is the space of uncertain parameters $\mathbb{D} \subseteq \mathbb{R}^n \times \mathbb{C}^m$ (e.g. a cylinder) topologically multiplied by the frequency spectrum Ω. L(d,s) $\in \mathbb{C}^{n \times n}$ is the stable loop transfer matrix. d is the vector of parametric uncertainty. The Nyquist template is $N = f(\mathbb{D} \times \Omega)$. The Horowitz template is $N_\omega = f(\mathbb{D}, \omega)$.

The least upper bound, k_M, such that the system is stable for all $d \in \mathbb{D}$ with $\|d\|_\infty < k_M$ is called the **robustness margin** of a control system. Intuitively, $k_M(j\omega)$ is achieved when the boundary of the template intercepts $0 + j0$. The natural question is what is the preimage of this situation in uncertainty space.

Now, the robustness margin problem with real parametric uncertainties becomes that of finding the smallest $\| d \|_\infty$ such that the template hits $(0 + j0)$. Structured singular value, μ, introduced in [12], is an equivalent way to formulate this problem where $\mu = 1/k_M$.

3 Mathematical Background

In this paper, we want to extend the previous results to propose a new method for the mixed k_M computation by using the following mathematical tools.

A **function** f of the complex variable z is **holomorphic** in an open set if it has a derivative at each point in that set [6]. A complex-valued function on a domain in $\mathbb{C}^n$ is holomorphic if an only if it is holomorphic in each variable separately [8]. The **open mapping theorem** says that a nonconstant holomorphic function is an open mapping [8].

A **critical point** q^* of a smooth map f_ω defined over $\mathbb{R}^n$ is a point where the rank of the Jacobian

$$J_{f_\omega} = \begin{pmatrix} \frac{\partial \Re f_\omega}{\partial q_1} & \frac{\partial \Re f_\omega}{\partial q_2} & \cdots & \frac{\partial \Re f_\omega}{\partial q_n} \\ \frac{\partial \Im f_\omega}{\partial q_1} & \frac{\partial \Im f_\omega}{\partial q_2} & \cdots & \frac{\partial \Im f_\omega}{\partial q_n} \end{pmatrix} \tag{1}$$

is < 2 [16, 19, 33, 22, 23].

The rank of the Jacobian may drop to either 1 or 0. A critical value is the value of the map $f_\omega(q^*)$ [20, 22].

The main difference between our method and the existing methods is using the polar form on a manifold instead of Euclidean coordinates for complex uncertainty[6, 8]. A manifold is defined as follows: An n-dimensional **manifold** is a Hausdorff space X with a countable basis such that each point x of X has a neighborhood that is homeomorphic with an open subset of $\mathbb{R}^n$ [33]. A stratification of an non-manifold is a decomposition of that into finitely manifolds. The manifolds in a stratification are called strata [15, 19, 20].

The **Sard's Theorem** says that if $f_\omega : X \longrightarrow Y$ is a smooth map between smooth manifolds, then almost every point in Y is a regular value of f_ω[16]. Let X be a metric space. As a subset N of $\mathbb{R}^n$, ∂N, the **boundary** of N , is defined as $\partial N = \overline{N} - Int(N)$. The closure of N, $\overline{N}$, is the intersection of all closed sets of $\mathbb{R}^n$ covering N, while the interior of N, $Int(N)$, is the union of all open sets contained in N. A collection $\mathcal{A}$ of subsets of a space X is said to **cover** X, or to be a **covering** of X, if the union of the elements of X if its elements are open subsets of X [33]. Hence, if Y is a subspace of X, a collection $\mathcal{A}$ of subsets of X is said to cover Y if the union of its elements contains Y.

4 Main Results

In this section, we will apply the topological concepts to propose a new method for the mixed stability robustness problem. Grid method has been criticized for its troublesome complexity. Therefore, instead of checking the template having passed through the origin, we will check the boundary of the template having passed through the origin.

One natural way to find $k_M(\omega)$ is to find the boundary of the template and determine the origin is inside the boundary curve or not. The stratified Morse theory has been applied to solve real k_m problem in [20]. For real uncertainties, the number of the set of equations need to be solved in order to obtain the critical points of the uncertainties depends on the number of the uncertainties. On the contrary, no matter how many uncertainties, for complex case, there is only **one set** of equations need be solved in order to obtain the critical points of the uncertainties. This has been shown in [28], which the stratified Morse theory also has been applied to solve the complex k_m problem. Here, we want to extend the previous results to solve mixed k_M problem. The basic idea for searching the singularity of the mixed uncertainty space by stratified Morse theory is as follows:

The usual uncertainty is closed. Hence, solid cylinder is not a manifold. A difficulty in dealing with the non-manifold, we prefer to resort to a more modern concept of stratified spaces [15, 19, 20, 24, 25]. Although the closed solid cylinder is not a manifold, it can nevertheless be decomposed into manifolds, i.e., solid cylinder = open cylinder ∪ open tube ∪ open disks . Each piece is a smooth manifold. Hence, we can use the Morse theory to find the singularities for each piece.

The boundary of a unit disc, $\partial\mathbf{D}$, is a unit circle, S^1. Due to Open Mapping Theorem or the lemma 1 of [12], the image of the distinguished boundary of complex uncertainties is first to intercept the origin [21]. We can use the **polar form** to express z as $z = ke^{j\phi}$ for fixed k [8, 25]. For example, in case of one complex uncertainty and one real uncertainty. The boundary of the uncertainties is a 2-dimensional manifolds, i.e., an open tube and open disks.

However, the boundary is still too big for us to compute k_M. Hence, we may ask the following question:
Can we narrow down the boundary to a smaller set ?
The answer is "Yes". The procedure to obtain the critical points of the open tube is as follows:
f_ω can be expressed as a function of q and ϕ, i.e.,

$$f_\omega^{q,\phi} = f_\omega(q,\phi)$$

The Jacobian becomes

$$J_{f_\omega^{q,\phi}} = \begin{pmatrix} \frac{\partial \Re f_\omega^{q,\phi}}{\partial q} & \frac{\partial \Re f_\omega^{q,\phi}}{\partial \phi} \\ \frac{\partial \Im f_\omega^{q,\phi}}{\partial q} & \frac{\partial \Im f_\omega^{q,\phi}}{\partial \phi} \end{pmatrix} \tag{2}$$

The constraints of the critical points are

$$\frac{\partial \Re f_\omega^{q,\phi}}{\partial q}\frac{\partial \Im f_\omega^{q,\phi}}{\partial \phi} - \frac{\partial \Im f_\omega^{q,\phi}}{\partial q}\frac{\partial \Re f_\omega^{q,\phi}}{\partial \phi} = 0 \tag{3}$$

Therefore, the number of constraints for critical points is equal to 1. There are 2 variables, i.e., q and ϕ. Hence, the solutions generically are curves.

For the general case, n real uncertainties and m complex uncertainties:

$$f_\omega^{q,\phi} = f_\omega(q_1, q_2, \cdots, q_n, \phi_1, \phi_2, \cdots, \phi_m)$$

The Jacobian becomes

$$J_{f_\omega^{q,\phi}} = \begin{pmatrix} \frac{\partial \Re f_\omega^{q,\phi}}{\partial q_1} & \frac{\partial \Re f_\omega^{q,\phi}}{\partial q_2} & \cdots & \frac{\partial \Re f_\omega^{q,\phi}}{\partial q_n} & \frac{\partial \Re f_\omega^{q,\phi}}{\partial \phi_1} & \frac{\partial \Re f_\omega^{q,\phi}}{\partial \phi_2} & \cdots & \frac{\partial \Re f_\omega^{q,\phi}}{\partial \phi_m} \\ \frac{\partial \Im f_\omega^{q,\phi}}{\partial q_1} & \frac{\partial \Im f_\omega^{q,\phi}}{\partial q_2} & \cdots & \frac{\partial \Im f_\omega^{q,\phi}}{\partial q_n} & \frac{\partial \Im f_\omega^{q,\phi}}{\partial \phi_1} & \frac{\partial \Im f_\omega^{q,\phi}}{\partial \phi_2} & \cdots & \frac{\partial \Im f_\omega^{q,\phi}}{\partial \phi_m} \end{pmatrix} \tag{4}$$

The constraints of singularities are

$$\left.\begin{array}{c} \frac{\partial \Re f_\omega^{q,\phi}}{\partial q_1}\frac{\partial \Im f_\omega^{q,\phi}}{\partial q_2} - \frac{\partial \Im f_\omega^{q,\phi}}{\partial q_1}\frac{\partial \Re f_\omega^{q,\phi}}{\partial q_2} = 0 \\ \cdots \\ \frac{\partial \Re f_\omega^{q,\phi}}{\partial q_1}\frac{\partial \Im f_\omega^{q,\phi}}{\partial q_n} - \frac{\partial \Im f_\omega^{q,\phi}}{\partial q_1}\frac{\partial \Re f_\omega^{q,\phi}}{\partial q_n} = 0 \\ \frac{\partial \Re f_\omega^{q,\phi}}{\partial q_1}\frac{\partial \Im f_\omega^{q,\phi}}{\partial \phi_1} - \frac{\partial \Im f_\omega^{q,\phi}}{\partial q_1}\frac{\partial \Re f_\omega^{q,\phi}}{\partial \phi_1} = 0 \\ \cdots \\ \frac{\partial \Re f_\omega^{q,\phi}}{\partial q_1}\frac{\partial \Im f_\omega^{q,\phi}}{\partial \phi_m} - \frac{\partial \Im f_\omega^{q,\phi}}{\partial q_1}\frac{\partial \Re f_\omega^{q,\phi}}{\partial \phi_m} = 0 \end{array}\right\} \tag{5}$$

Therefore, the number of constraints for critical points in the case of n real uncertainties and m complex uncertainties is equal to the number of the variables minus one ($n+m-1$). The solutions generically are curves. Similarly, we need to compute the critical points of $f_\omega^{q,\phi} = f_\omega(q_2, \cdots, q_n, \phi_1, \phi_2, \cdots, \phi_m)$ for fixed q_1 and so forth. Therefore, in mixed case, it need to compute more than one set of equations in order to obtain the critical points. This situation is similar to all real uncertainties case. On the contrary, for complex uncertainty only, we can obtain the critical points by solving one set of equations. However, we can use eq. (5) and similar equations to find all the critical points. Once all the critical points are found, we can easily to determine the boundary of the Horowitz template cover the origin or not. Hence, we need solve systems of equations in order to obtain the boundary of the Horowitz template. It is well known from Galois theory that even single univariate equations of degree 5 or higher cannot (generally) be explicitly solve by radicals [14]. Eq. (5) involves multivariables and higher degree. Hence, to find a general formula to solve eq. (5) is almost no hope. Our goal will be to find all approximate of solutions. How to solve eq. (5) is related to Algebraic Geometry and Groebner basis [7]. See [7, 14] for detail. However, the similar problems have been formulated in other applications, i.e., robotics, computer graphics, molecular modeling [7]. Therefore, this is an interesting topic for applications. In addition, we can use the Symbolic tools to find the solutions [14]. Due to this method using the boundary of the template to compute k_m, it is called "Boundary Approach" [24, 25].

5 Symbolic Computation

Symbolic computation, also known as Computer Algebra, is defined as computation with symbols representing mathematical objects [9, 14, 18]. Symbolic computation systems are powerful tools for mathematicians, engineers, ... to compute not only with numbers, but also with symbols, formulae, equations, and so on.

Based on earlier discussion, it is hard to implement the Boundary Approach to compute k_M under the pure numeric computing environment. The reasons are as follows:

1. It need partial differentiate the Nyquist map. When we use the definition of critical point to calculate singularities, we need to partial differentiate the Nyquist map relative to uncertainties ϕ_i as in eq. (4).
2. It need solve the set of equations in eq. (5).
3. It need change variables.

It is tedious to obtain the critical curves of the manifold by pure numeric computing environment. We need to solve a set of equations in order to find the critical curves. If the robustness margin problem has more uncertainties, it is hard to obtain the critical curves. Therefore, symbolic computation systems are needed for developing a software package to compute the stability robustness margin using the topological approach.

Fortunely, the symbolic computing has matured to a point that it is easy to be used and several software packages such as MACSYMA, REDUCE, MATLAB, Maple, Mathematica, etc., all include symbolic computation tools [14, 18]. Moreover, symbolic computation tools has been successful used to solve some control problems by several researchers [11]. Hence, we will use these tools to provide an efficient way for solving the stability robustness problem. Due to the reformulation of the k_M problem as to solve eq.

(5), the function "solve" in the symbolic computation tools is the most critical procedure for k_M computation.

6 An Algorithm for Stability Margin Computation

Now, we apply the results of the previous sections to obtain an algorithm for k_M computation without suffering the same troublesome complexity of the grid method as follows:

1. Initialization.
2. Use the symbolic computation tools to solve eq. (5) for critical points.
3. Use the Nyquist map to find the boundary of the Horowitz template.
4. If the boundary of the template intercept the origin, $k_M = k$. If the boundary of the template does not intercept the origin, increase k and go to step 2.
5. If ω is bigger enough, stop computing (the robustness margin is obtain). Otherwise, increase ω and go to 2.

In step 4, if the origin is covered by the open covering of the boundary of the template, we think that the boundary of the template intercept the origin [24]. In step 4, we should avoid increasing k too quickly [24].

7 Examples

The algorithm was implemented in Matlab with the symbolic Math Toolbox and applied to the following examples and was found it performed satisfactory [17].

Example 1 *Consider the following example:*

$$f_\omega = z * q + z + q + 1 \tag{6}$$

The Jacobian of eq. (6) is

$$J_{f^\phi_\omega} = \begin{pmatrix} cos\phi + 1 & -qsin\phi - sin\phi \\ sin\phi & qcos\phi + cos\phi \end{pmatrix}$$

The constraint of the critical points is $(q+1)(cos\phi + 1) = 0$. *Hence,* $\phi = \pi$ *or* $q = -1$. *The simulation results are shown in Figure 1 and Figure 2 for the Horowitz template and the boundary, respectively. The boundary of the template is a curve.*

Example 2 *Consider the following example:*

$$f_\omega = z_1 * z_2 + z_1 + 1 \tag{7}$$

The Jacobian of eq. (7) is

$$J = \begin{pmatrix} -sin(\phi_1 + \phi_2) - sin\phi_1 & -sin(\phi_1 + \phi_2) \\ cos(\phi_1 + \phi_2) + cos\phi_1 & cos(\phi_1 + \phi_2) \end{pmatrix}$$

The constraint of the critical points is $-sin\phi_1 * cos(\phi_1 + \phi_2) - sin(\phi_1 + \phi_2) * cos\phi_1 = 0$ *Hence,* $\phi_2 = 0$ *or* $\phi_2 = \pi$. *The Horowitz template is shown in Fig 3. The boundary of the template is shown in Figure 4.*

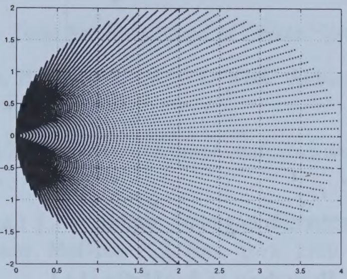


Figure 1: f=z*q+z+q+1

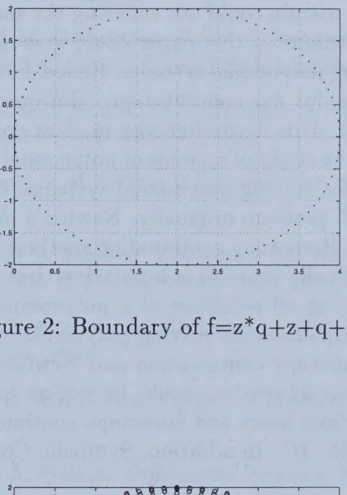


Figure 2: Boundary of f=z*q+z+q+1

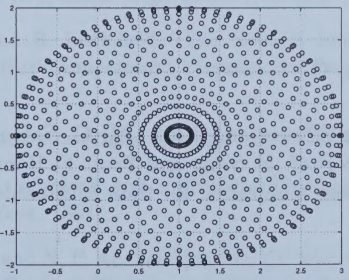


Figure 3: f=z1*z2+z1+1

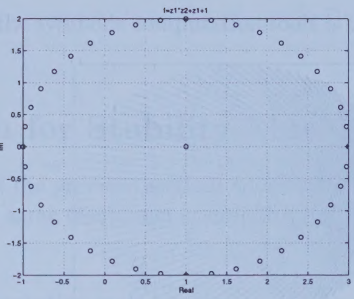


Figure 4: Boundary of f=z1*z2+z1+1

8 Conclusion

In this paper, we have proposed a new method for the mixed robustness margin computation. This method has reformulated the mixed k_M problem as to solve polynomial systems. The solutions are critical points of manifolds. Once the critical points are found, obtaining the accurate robustness margin could not suffering the same troublesome complexity of the grid method. The results of this paper could give us a new direction for mixed k_M computation, i.e. solving polynomial systems. Hence, how to solve polynomial systems becomes the central issue for k_M computation. However, solving polynomial systems is NP-hard in general [13]. This is another way to show complex k_M problem is NP-hard. Algebraic Geometry is the study of systems of polynomial equations. Groebner bases is the well-known method for solving polynomial systems [7]. However, in practice, it works well only for "small" problem originally. Newton's method is an iteration algorithm good for local analysis. Homotopy continuation method is another method to study polynomial systems. It originally requires computation time proportional to total degree of polynomials for computing all solutions of a polynomial system. Fortunely, motivated by the need of fast compuation for solving polynomial systems, a lot of new results about Groebner bases, homotopy continuation and Newton's method have been obtained recently. Solving polynomial systems could be not as hard as it used to be. Hence, Algebraic Geometry, Groebner bases and homotopy continuation method will be major tools for our approach [7, 30, 31]. In addition, Symbolic Computation Tools will be a key to this method.

Acknowledgement

I am grateful to Prof. E. A. Jonckheere and Prof. M. G. Safonov for many helpful discussions. Especial thanks go to Prof. P. Yang and Prof. F. Bonahon, Department of Mathematics, University of Southern California, for their invaluable suggestions.

References

[1] J. Ackermann, A. Bartlett, D. Kaesbauer, W. Sienel, and R. Steinhauser. *Robust Control-Systems with Uncertain Physical Parameters.* Springer-Verlag, New York, 1993.

[2] V. I. Arnold, S. M. Gusein-Zade, and A. N. Varchenko. *Singularities of Differentiable Maps.* Birkhäuser, Boston, 1988.

[3] B. Barmish. *New Tools for Robustness of Linear Systems.* Macmillan, New York, 1994.

[4] S. Bhattacharyya, H. Chapellat, and L. Keel. *Robust Control-The Parametric Approach.* Prentice-Hall, New York, 1995.

[5] J. W. Bruce and P. J. Giblin. *Curves and Singularities.* Cambridge University Press, Cambridge, 1992.

[6] R. V. Churchill and J. W. Brown. *Complex Variables and Applications.* Addison-Wesley Publishing Company, McGraw Hill, New York, 1990.

[7] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms.* Springer-Verlag, New York, 1992.

[8] J. P. D'Angelo. *Several Complex Variables and the Geometry of Real Hypersurfaces.* CRC Press, Inc., Boca Raton, 1993.

[9] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation.* Academic Press, New York, 1988.

[10] R. de Gaston and M. G. Safonov. Exact Calculation of the Multiloop Stability Margin. *IEEE Trans. Automat. Contr.*, AC-33:156–171, 1988.

[11] B. de Jager. The Use of Symbolic Computation in Nonlinear Control: Is it Viable? *IEEE Trans. Automat. Contr.*, AC-40:84–89, 1995.

[12] J. C. Doyle. Analysis of Feedback Systems with Structured Uncertainty. *IEE Proc.*, 129:242–250, 1982.

[13] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP Completeness.* W. H. Freeman, New York, 1979.

[14] K. O. Geddes, S. R. Czapor, and G. Labahn. *Algorithms for Computer Algebra.* Kluwer Academic Publishers, 1992.

[15] M. Goresky and R. MacPheron. *Stratified Morse Theory.* Springer-Verlag, New York, 1987.

[16] V. Guillemin and A. Pollack. *Differential Topology.* Prentice-Hall, New Jersey, 1974.

[17] D. Hanselman and B. Littlefield. *The Student Edition of MATLAB–The Language of Technical Computing–Computation Visualization Programming.* Prentice-Hall, New Jersey, 1997.

[18] A. Heck. *Introduction to Maple.* Springer-Verger, New York, 1993.

[19] E. A. Jonckheere. *Algebraic and Differential Topology of Robust Stability.* Oxford University Press, New York, 1997.

[20] E. A. Jonckheere and N. P. Ke. Stability Margin relative to Stratified Uncertainty Space. In *Proc. 34th IEEE Conf. on Decision and Control*, pages 1322–1323, New Orleans, Louisiana, 1995.

[21] E. A. Jonckheere and N. P. Ke. Complex-Analytic Theory of the μ-Function. In *Proc. American Control Conf.*, pages 366–371, Albuquerque, New Mexico, 1997.

[22] E. A. Jonckheere and N. P. Ke. Real versus Complex Robustness Margin Continuity as a Smooth versus Holomorphic Singularity Problem. In *Proc. 36th IEEE Conf. on Decision and Control*, pages 3266–3271, San Diego, California, 1997.

[23] N. P. Ke. Topological concepts for robustness margin problems. In *Proc. of the 35th Annual Allerton Conf. on Communication Control and Computing*, pages 878–887, Monticello, Illinois, 1997.

[24] N. P. Ke. An Algorithm of Symbolic Computation for Stability Margin. In *Proc. American Control Conf.*, pages 2304–2308, Philadelphia, Pennsylvania, 1998.

[25] N. P. Ke. Singularity Theory and Stability Robustness Analysis. In *Proc. IEE CONTROL'98*, pages 624–629, University of Wales, Swansea, UK, 1998.

[26] N. P. Ke. *Stability Robustness Analysis - A Differential Topological Approach.* PhD thesis, University of Southern California, Los Angeles, California, May 1998.

[27] N. P. Ke. Value Set versus Horowitz Template in Robust Stability Analysis. In *Proc. 37th IEEE Conf. on Decision and Control*, Tampa, Florida, 1998. to appear.

[28] N. P. Ke. A New Method to Compute Complex Robustness Margin. In *IFAC'99*, Beijing, P. R. China, 1999. summitted.

[29] F. J. Kraus, B. D. O. Anderson, and M. Mansour. Robust Stability of Polynomials with Multilinear parameter Dependence. *Int. J. Contr.*, 50:1745–1762, 1989.

[30] T. Y. Li. Numerical Solution of Multivariate Polynomial Systems by Homotopy Continuation Methods. *Acta Numerica*, pages 399–436, 1997.

[31] D. Manocha. Computing Selected Solutions of Polynomial Equations. In *ISAAC'94*, pages 1–8, Oxford, UK, 1994.

[32] M. Morse. Relations between the critical points of a real function of n independent variables. *Trans. American Math. Soc.*, 27:345–396, 1925.

[33] J. R. Munkres. *Topology - a first course.* Prentice-Hall, New Jersey, 1975.

[34] S. Poljak and J. Rohn. Checking Robust Nonsingularity is NP-hard. *Math. Contr. Signals Systems*, 6:1–9, 1993.

[35] M. Safonov and M. Athans. A Multiloop Generalization of the Circle Criterion for Stability Margin Analysis. *IEEE Trans. Automat. Contr.*, AC-26:415–422, 1981.

[36] O. Toker and H. Ozbay. On the Complexity of Purely Complex μ Computation and Related Problems in Multidimensional Systems. *IEEE Trans. Automat. Contr.*, AC-43:409–414, 1998.

State Feedback Stabilization Over a Finite Time Interval of Linear Systems Subject to Norm Bounded Uncertainties

F. Amato, M. Ariola
Dipartimento di Informatica e Sistemistica
Università degli Studi di Napoli Federico II
Via Claudio 21, Napoli, 80125
ITALY
{amato,ariola}@disna.dis.unina.it

P. Dorato
EECE Department
The University of New Mexico
Albuquerque, NM 87131
USA
peter@eece.unm.edu

Abstract

In this paper we consider the state feedback, finite-time robust stabilization problem for linear systems subject to time-varying norm-bounded uncertainties. We provide a sufficient condition for state feedback finite-time stabilization in terms of the solution of a *Generalized Eigenvalue Problem*. The proposed technique is illustrated via a detailed numerical example.

1 Introduction

In the last years a big effort has been spent to study the robust stability problem for linear systems subject to uncertainties (see, for instance, the tutorial books [4] and [5]). The work of control scientists and engineers has mainly focused on robust *Lyapunov Stability*; as it is well known, Lyapunov stability deals with the steady-state behaviour of linear systems, because it looks at the asymptotic pattern of system trajectories.

In many practical applications, however, also the transient behaviour of the system plays an important rôle; for example, due to the presence of some saturation elements in the control loop, it could be required that the trajectories of the controlled system do not exceed given bounds. This kind of specifications are well captured by the concept of Finite-Time Stability [3], whose exact definition is provided in Section 2.

In this paper we focus our attention on linear, time-varying systems, subject to norm bounded uncertainties. For this class of uncertain systems, in Section 3 we provide a sufficient condition for Finite-Time Stabilization via state-feedback in terms of the solvability of a differential matrix inequality. Then in Section 4 we show how for the *time-invariant* case this condition can be reduced to an optimization problem involving the feasibility of a system of LMIs [2]. Some numerical examples, shown in Section 5, highlight the main points discussed in the paper. Finally, some concluding remarks and plans for future research, given in Section 6, end the paper.

2 Problem Statement

The general idea of *finite-time stability* concerns the boundedness of the state of a system over a finite time interval. This concept can be formalized through the following definition, which is an extension of the one given in [3].

Definition 1 The time-varying linear system

$$\dot{x} = A(t)x \qquad t \in [0, +\infty) \tag{1}$$

is said to be Finite-Time Stable (FTS) with respect to the given quadruple (c_1, c_2, T, R), with $c_2 > c_1$ and $R > 0$ if

$$x^T(0)Rx(0) \le c_1 \Rightarrow x^T(t)Rx(t) < c_2 \qquad \forall t \in [0, T] \tag{2}$$

■

Remark 1 *Relation between Lyapunov Stability and FT Stability*
Lyapunov Asymptotic Stability and FT Stability are independent concepts: a system which is FT stable could be not Lyapunov asymptotically stable, whereas a Lyapunov asymptotically stable system could be not FT stable if, during the transient, its state exceeds the prescribed bounds. ■

In the following we shall denote by $\mathcal{PC}$ ($\mathcal{PC}^1$) the space of uniformly bounded, piecewise continous (piecewise continuosly differentiable), real matrix-valued functions defined on $[0, +\infty)$.

Let us consider the following linear, time-varying system subject to time-varying uncertainties

$$\dot{x}(t) = [A(t) + \Delta A(t)]\, x(t) + [B(t) + \Delta B(t)]\, u(t) \tag{3}$$

with A and B belonging to $\mathcal{PC}$. We assume that the uncertain part in (3) is in the so-called *structured, one block form*

$$[\Delta A(t) \quad \Delta B(t)] = F(t)\Delta(t)\, [E_1(t) \quad E_2(t)] \tag{4}$$

where F, E_1 and E_2 belongs to $\mathcal{PC}$ and the unknown, real matrix-valued function Δ belong to the class

$$\mathcal{D} := \left\{\Delta : [0, +\infty) \mapsto \mathbb{R}^{i \times j} \,|\, \Delta \text{ is Lebesgue measurable, } \Delta(t)\Delta(t)^T \le I\right\}. \tag{5}$$

Concerning system (3), the goal of this paper is that of solving the following problems.

Problem 1 Given system (3) and the quadruple (c_1, c_2, T, R), find a state feedback controller

$$u = K(t)x\,, \qquad K \in \mathcal{PC} \tag{6}$$

such that the closed loop system given by the connection of (3) with (6) is FTS with respect to (c_1, c_2, T, R), for all $\Delta \in \mathcal{D}$. ■

Problem 2 Given system (3) and the triplet (c_1, c_2, R), find a state feedback controller in the form (6) which maximizes the positive number T, and renders the closed loop system given by the connection of (3) with (6) FTS with respect to (c_1, c_2, T, R), for all $\Delta \in \mathcal{D}$. ■

Problem 3 Given system (3) and the triplet (c_1, T, R), find a state feedback controller in the form (6) which minimizes the positive number c_2 $(c_2 > c_1)$, and renders the closed loop system given by the connection of (3) with (6) FTS with respect to (c_1, c_2, T, R), for all $\Delta \in \mathcal{D}$. ■

We are interested in the solution of Problem 1, because the solution of Problem 2 (3) can be obtained via a trivial binary search algorithm over T (c_2) based on the solution of Problem 1.

3 Main Result

In this section we shall provide a sufficient condition for the solvability of Problem 1; to state this condition we need the following preliminary result. Note that in the sequel, for the sake of brevity, we shall often omit the time argument.

Let us consider system (3) with $u(t) = 0$

$$\dot{x}(t) = [A(t) + F\Delta E_1]\, x(t) =: \hat{A}(t)x(t) \tag{7}$$

Lemma 1 *System (7) is FTS with respect to (c_1, c_2, T, R) for all $\Delta \in \mathcal{D}$ if, letting $\tilde{Q} = R^{-1/2}QR^{-1/2}$, there exist a nonnegative scalar α, a symmetric positive definite matrix function $Q \in \mathcal{PC}^1$ and a positive definite scalar function $\mu \in \mathcal{PC}$ such that for all $t \in [0, T]$*

$$-\dot{\tilde{Q}} + A\tilde{Q} + \tilde{Q}A^T + \frac{1}{\mu^2}FF^T + \mu^2\tilde{Q}E_1^T E_1\tilde{Q} - \alpha\tilde{Q} \ < \ 0 \tag{8a}$$

$$cond(Q) \ < \ \frac{c_2}{c_1}e^{-\alpha T} \tag{8b}$$

Where $cond(\cdot)$ denotes the condition number of the argument.

Proof. Assume that there exist an $\alpha \geq 0$ and a symmetric matrix $Q > 0$ such that for all $t \in [0, T]$ inequalities (8) are satisfied. Using the following inequality

$$\frac{1}{\mu^2}FF^T + \mu^2\tilde{Q}E_1^T E_1\tilde{Q} \geq \frac{1}{\mu^2}F\Delta\Delta^T F^T + \mu^2\tilde{Q}E_1^T E_1\tilde{Q} \geq F\Delta E_1\tilde{Q} + \tilde{Q}E_1^T\Delta^T F^T \tag{9}$$

from (8a) we easily have

$$-\dot{\tilde{Q}} + \hat{A}\tilde{Q} + \tilde{Q}\hat{A}^T - \alpha\tilde{Q} < 0 \tag{10}$$

where $\hat{A}$ is defined in (7). Pre and post-multiplying (10) by $\tilde{Q}^{-1}$ we obtain the equivalent condition

$$-\tilde{Q}^{-1}\dot{\tilde{Q}}\tilde{Q}^{-1} + \tilde{Q}^{-1}\hat{A} + \hat{A}^T\tilde{Q}^{-1} < \alpha\tilde{Q}^{-1} \tag{11}$$

Now let $V(x) = x^T\tilde{Q}^{-1}x$. Denoting, as usual, by $\dot{V}$ the derivative of V along the solution of system (7) and since $\frac{d}{dt}\tilde{Q}^{-1} = -\tilde{Q}^{-1}\dot{\tilde{Q}}\tilde{Q}^{-1}$, from (11) we have

$$\frac{\dot{V}(x(t))}{V(x(t))} < \alpha \tag{12}$$

From (12) follows

$$x^T(t)\tilde{Q}^{-1}x(t) < x^T(0)\tilde{Q}^{-1}x(0)e^{\alpha t} \tag{13}$$

Now we have

$$\begin{aligned} x^T(t)\tilde{Q}^{-1}x(t) &\geq \lambda_{min}(Q^{-1})x^T(t)Rx(t) && \text{(14a)} \\ x^T(0)\tilde{Q}^{-1}x(0)e^{-|\alpha|t} &\leq \lambda_{max}(Q^{-1})x^T(0)Rx(0)e^{\alpha t} \\ &\leq \lambda_{max}(Q^{-1})c_1 e^{\alpha T} && \text{(14b)} \end{aligned}$$

Putting together (13) and (14) we have

$$\begin{aligned} x^T(t)Rx(t) &< cond(Q^{-1})c_1 e^{\alpha T} \\ &= cond(Q)c_1 e^{\alpha T} && \text{(15)} \end{aligned}$$

From (15) readily follows that (8b) implies, for all $t \in [0,T]$, $x^T(t)Rx(t) < c_2$; from the last consideration the proof follows. ■

Remark 2 If $\alpha = 0$, inequality (8b) is independent of T. In this case, if inequalities (8) hold for all $t \in [0,+\infty)$, system (7) is FTS with respect to (c_1, c_2, T, R) for all $T \in (0,+\infty)$; it is easy to show that if Lemma 1 is satisfied for $\alpha = 0$ and for all $t \in [0,+\infty)$, FT Stability *implies* Quadratic Stability which in turn implies Lyapunov asymptotic stability for all $\varDelta \in \mathcal{D}$ [1] (see also Remark 1). ■

Now let us go back to the original problems concerning system (3).

The following theorem gives a sufficient condition for the finite-time stabilizability of system (3) via state feedback. The closed loop system given by the connection of (3) with (6) has the form

$$\dot{x}(t) = [A(t) + B(t)K(t) + F(t)\varDelta(t)(E_1(t) + E_2(t)K(t))]\,x(t) \tag{16}$$

Theorem 1 *Problem 1 is solvable if there exist a nonnegative scalar α, a symmetric positive definite matrix function $Q \in \mathcal{PC}^1$, a matrix $L \in \mathcal{PC}$ and a positive definite scalar function $\mu \in \mathcal{PC}$ such that for all $t \in [0,T]$*

$$\begin{aligned} &-\dot{\tilde{Q}} + A\tilde{Q} + \tilde{Q}A^T + BL + L^TB^T + \frac{1}{\mu^2}FF^T + \mu^2\tilde{Q}E_1^TE_1\tilde{Q} + \mu^2L^TE_2^TE_2L \\ &+\mu^2\tilde{Q}E_1^TE_2L + \mu^2L^TE_2^TE_1\tilde{Q} - \alpha\tilde{Q} < 0 \\ &cond(Q) < \frac{c_2}{c_1}e^{-\alpha T} \end{aligned}$$

In this case a controller which solves Problem 1 is given by $K = L\tilde{Q}^{-1}$.

Proof. The proof easily follows from Lemma 1, replacing A with $A + BK$ and E with $E_1 + E_2K$ (compare (7) with (16)), and letting, according to [7], $L = K\tilde{Q}$. ■

Remark 3 If Theorem 1 holds for $\alpha = 0$ and for all $t \in [0,+\infty)$, the closed loop system (16) is quadratically stable and therefore it is uniformly asymptotically stable on the infinite interval $[0,+\infty)$ (see also Remarks 1 and 2). ■

4 The Time-Invariant Case

In this section we focus on systems in the form (3), where the *certain* matrices are *time-invariant*

$$\dot{x}(t) = [A + \Delta A(t)]\, x(t) + [B + \Delta B(t)]\, u(t) \tag{17}$$

with

$$[\Delta A(t) \quad \Delta B(t)] = F\Delta(t)\, [E_1 \quad E_2] \tag{18}$$

For this class of systems we shall provide a sufficient condition for *time-invariant* state feedback FT Stabilization which can be reduced to an optimization problem involving Linear Matrix Inequalities (LMIs) [2].

The following result is a direct consequence of Theorem 1 for time-invariant systems.

Theorem 2 *Problem 1 (stated for system (17)) is solvable if there exist a nonnegative scalar α, a symmetric positive definite matrix Q and a matrix L such that*

$$\begin{aligned} &A\tilde{Q} + \tilde{Q}A^T + BL + L^T B^T + FF^T + \tilde{Q}E_1^T E_1 \tilde{Q} + L^T E_2^T E_2 L \\ &+\tilde{Q}E_1^T E_2 L + L^T E_2^T E_1 \tilde{Q} - \alpha\tilde{Q} < 0 \\ &cond(Q) < \frac{c_2}{c_1} e^{-\alpha T} \end{aligned}$$

In this case a (time-invariant) controller which solves Problem 1 is given by $K = L\tilde{Q}^{-1}$.

Proof. Straightforward. ■

Remark 4 In Theorem 2 we restricted our attention to *constant* matrix functions Q, L and to a constant scalar μ in order to get conditions involving LMIs. Furthermore we let $\mu = 1$ without any loss of generality. ■

For a given α, the feasibility of the conditions stated in Theorem 2 can be turned into an LMIs based optimization problem. Indeed these conditions can be equivalently restated as the following *Generalized Eigenvalue Problem* (GEVP) [2].

Generalized Eigenvalue Problem

Given system (17) and the quadruple (c_1, c_2, T, R), fix an $\alpha \geq 0$ and solve

$$\min_{L,Q} \gamma$$

$$s.t.$$

$$\begin{pmatrix} A\tilde{Q} + \tilde{Q}A^T + BL + L^T B^T + FF^T - \alpha\tilde{Q} & \left(E_1\tilde{Q} + E_2 L\right)^T \\ E_1\tilde{Q} + E_2 L & -I \end{pmatrix} < 0$$

$$I < Q < \gamma I$$

If $\gamma < \frac{c_2}{c_1} e^{-\alpha T}$, system (17) is FTS with respect to (c_1, c_2, T, R). ■

Obviously a binary search is necessary to find a feasible α.

5 Numerical Examples

In this section we will show three different examples where we applied the optimization procedure stated in the previous section. In all three cases, we solved Problem 2 (see Section 2) with the aid of the LMI Toolbox [6].

Example 1
Let us consider system (17) with

$$A = \begin{pmatrix} 0 & 20 \\ -1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

$$F = \begin{pmatrix} -1 & 0 \\ 1 & -2 \end{pmatrix}, \quad E_1 = \begin{pmatrix} 0 & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad E_2 = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}$$

Moreover let $c_1 = 1$, $c_2 = 100$ and $R = I$.

We got $T_{max} = 0.58$ for $\alpha = 3.35$ with the controller

$$K = \begin{pmatrix} 0.29 & 3.01 \end{pmatrix}$$

■

Example 2
Let us consider system (17) with

$$A = \begin{pmatrix} 0 & 1 \\ 0 & \frac{1}{10} \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$F = \begin{pmatrix} \frac{1}{10} & \frac{1}{10} \\ 0 & \frac{1}{10} \end{pmatrix}, \quad E_1 = \begin{pmatrix} \frac{1}{10} & 0 \\ \frac{1}{10} & \frac{1}{10} \end{pmatrix}, \quad E_2 = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}$$

Moreover let $c_1 = 1$, $c_2 = 10$ and $R = I$. Note that in this case *the nominal system* (A, B) *is not stabilizable*; therefore FT Stability is all we can look for. We got $T_{max} = 5.16$ for $\alpha = 0.45$ with the controller

$$K = \begin{pmatrix} -4.25 & 0.01 \end{pmatrix}$$

■

Example 3
Let us consider system (17) with

$$A = \begin{pmatrix} 0 & 1 \\ 0 & \frac{1}{10} \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$F = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad E_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad E_2 = \begin{pmatrix} \frac{1}{10} \\ 0 \end{pmatrix}$$

We got the controller

$$K = \begin{pmatrix} -109.8 & -96.1 \end{pmatrix}$$

for $\alpha = 0$, with a Q matrix with condition number *cond*$(Q) = 1.04$. Therefore in this case the closed loop system is FTS with respect to (c_1, c_2, T, I), with $c_1 = 1$ and $c_2 = 1.04$ for all $T \in (0, +\infty)$ and *is also quadratically stable* (see Remarks 2 and 3). ■

6 Concluding Remarks

In this paper we have considered the Finite-Time Stabilization problem via state-feedback for linear, time-varying systems subject to norm bounded uncertainties. A sufficient condition guaranteeing Finite-Time Stabilization, based on the solution of a differential matrix inequality, has been provided; in the *time-invariant* case, such condition reduces to an optimization problem involving the feasibility of a system of LMIs. To put into evidence the main points, some detailed numerical examples have been included.

Future work will be devoted to cope with the following problems

- Extension to discrete-time systems
- Dynamic Output feedback control (when the state is not available for feedback).

References

[1] Barmish B. R., "Stabilization of uncertain systems via linear control," *IEEE Trans. Auto. Control*, **28**, pp. 848–850, Aug. 1983.

[2] Boyd S., L. El Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Press, 1994.

[3] Dorato P., "Short time stability in linear time-varying systems," *Proc. IRE International Convention Record Part 4*, pp. 83–87, 1961.

[4] Dorato P. (Ed.), *Robust Control*, IEEE Press, New York, 1987.

[5] Dorato P. (Ed.), *Recent Advances in Robust Control*, IEEE Press, New York, 1990.

[6] Gahinet P., A. Nemirovski, A. J. Laub and M. Chilali, *LMI Control Toolbox*, the Mathworks Inc, Natick, MA, 1995.

[7] Geromel J. C., P. L. D. Peres and J. Bernussou, " On a convex parameter space method for linear control design of uncertain systems," *SIAM J. Contr. Opt.*, **29**, 381–402, 1991.

Iterative Signature Adaptation for Capacity Maximization of CDMA Systems*

Sennur Ulukus
AT&T Labs-Research
ulukus@research.att.com

Roy D. Yates
WINLAB, Rutgers University
ryates@winlab.rutgers.edu

Abstract

For single cell synchronous code division multiple access (CDMA) systems, both the information theoretic capacity and the network capacity have been identified. In both cases, it was shown that if the number of users, N, is no more than the processing gain, L, then orthogonal signatures are optimal while if $N > L$, then signatures which satisfy the Welch bound on the total squared correlation with equality (called WBE sequences) are optimal. This paper presents an algorithm which iteratively updates the signatures in a distributed fashion, starting from an initial set of signatures. Under mild conditions on the initial set of signatures, we prove that the algorithm converges to a set of orthogonal signatures if $N \leq L$ and to a WBE set if $N > L$. At each step, the algorithm replaces one signature from the set with the normalized linear MMSE receiver filter corresponding to that signature. Since the MMSE filter can be obtained by a distributed algorithm for each user, the algorithm is amenable to distributed implementation.

1 Introduction

In CDMA systems, users modulate their information streams with high frequency waveforms called CDMA codes or signature sequences. The signature waveform of user i, denoted by $s_i(t)$, is non-zero only in the bit interval $[0, T_b]$ and is normalized to unit energy, i.e., $\int_0^{T_b} s_i^2(t)dt = 1$. The received base band signal at the base station is

$$r(t) = \sum_{i=1}^{N} \sqrt{p_i} b_i s_i(t) + n(t) \tag{1}$$

where p_i and b_i are the received power and the information bit of user i, and $n(t)$ is the additive white Gaussian noise (AWGN) with zero mean and power spectral density σ^2.

We define the chip waveform to be $\psi(t)$, $t \in [0, T_c]$ and 0 elsewhere, where T_c is the chip duration. Thus $\{\psi(t - iT_c), i = 0, \ldots, L-1\}$, where $L = T_b/T_c$ is the processing gain, is a basis for the signal space. This allows us to represent the signature sequences of the users with L dimensional vectors. We will use $\boldsymbol{s}_i$ to denote the signature sequence of user i. We define an $L \times N$ matrix $\boldsymbol{S}$ which contains the signature sequences of all users in its columns, i.e., $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N]$ and a $N \times N$ diagonal matrix $\boldsymbol{P}$ whose ith diagonal element is p_i. In terms of signal vectors, the received signal can be written as

$$\boldsymbol{r} = \sum_{i=1}^{N} \sqrt{p_i} b_i \boldsymbol{s}_i + \boldsymbol{n} \tag{2}$$

where $\boldsymbol{n}$ is a Gaussian random vector with $E\big[\boldsymbol{n}\boldsymbol{n}^\top\big] = \sigma^2 \boldsymbol{I}_L$, where $\boldsymbol{I}_L$ is the $L \times L$ identity matrix.

---

*Supported in part by NSF Grant NCR 95-06505.

The capacity of a single cell synchronous CDMA system has been studied in two different contexts. In [1,2], the information theoretic capacity region of a CDMA channel is

$$\mathcal{C} = \bigcap_{J \subseteq \{1,\cdots,N\}} \left\{ (R_1, \cdots, R_N) : 0 \le \sum_{j \in J} R_j \le \frac{1}{2} \log \left[\det \left(\boldsymbol{I}_L + \sigma^{-2} \boldsymbol{S}_J \boldsymbol{P}_J \boldsymbol{S}_J^\top \right) \right] \right\} \tag{3}$$

where $\boldsymbol{S}_J$ is the $L \times |J|$ matrix which is obtained from $\boldsymbol{S}$ by striking out the columns of $\boldsymbol{S}$ whose indices do not belong to the subset J, and $\boldsymbol{P}_J$ is a $|J| \times |J|$ diagonal matrix which is obtained from $\boldsymbol{P}$ by striking out the rows and columns of $\boldsymbol{P}$ whose indices do not belong to the subset J.

An important measure of overall information capacity of a multiaccess channel is the *sum* capacity

$$C_{\text{sum}} = \max_{(R_1,\cdots,R_N) \in \mathcal{C}} \sum_{i=1}^{N} R_i \tag{4}$$

From (3), the sum capacity of a CDMA channel is

$$C_{\text{sum}} = \frac{1}{2} \log \left[\det \left(\boldsymbol{I}_L + \sigma^{-2} \boldsymbol{S} \boldsymbol{P} \boldsymbol{S}^\top \right) \right] \tag{5}$$

When the powers of the users are the same, $p_i = p$ for all i, (5) reduces to

$$C_{\text{sum}} = \frac{1}{2} \log \left[\det \left(\boldsymbol{I}_L + \frac{p}{\sigma^2} \boldsymbol{S} \boldsymbol{S}^\top \right) \right] = \frac{1}{2} \log \left[\det \left(\boldsymbol{I}_N + \frac{p}{\sigma^2} \boldsymbol{S}^\top \boldsymbol{S} \right) \right] \tag{6}$$

where the last equality follows from the fact that for any two matrices $\boldsymbol{A}_{K \times M}$ and $\boldsymbol{B}_{M \times K}$,

$$\det \left(\boldsymbol{I}_K + \boldsymbol{A}\boldsymbol{B} \right) = \det \left(\boldsymbol{I}_M + \boldsymbol{B}\boldsymbol{A} \right) \tag{7}$$

Clearly, the sum capacity depends on the signature sequence set being used. It was shown in [2] that the sum capacity for equal received powers is maximized if the signature sequences are chosen such that if $N \le L$,

$$\boldsymbol{S}^\top \boldsymbol{S} = \boldsymbol{I}_N \tag{8}$$

and if $N > L$,

$$\boldsymbol{S}\boldsymbol{S}^\top = \frac{N}{L} \boldsymbol{I}_L \tag{9}$$

The signature sequence sets satisfying (8) contain N orthogonal signature sequences in L dimensional vector space. When $N \le L$, N such sequences can be found. The sequence sets satisfying (9) are named Welch Bound Equality (WBE) sequence sets in [2] because they satisfy the Welch's bound on the sum of the squares of the cross correlations of unit energy sequences with equality.

In [3], the *network* capacity of a CDMA system is defined in terms of the maximum number of admissible users. Given the processing gain L and a common Signal to Interference Ratio (SIR) target β, N users are said to be admissible if there exist positive powers p_i and signature sequences $\boldsymbol{s}_i$ such that each user has an SIR at least as large as the target SIR β. The network capacity was found for two kinds of linear receiver structures in [3]: matched filters and Minimum Mean Squared Error (MMSE) filters [4,5]. The MMSE filter for the kth user, $\boldsymbol{c}_k$, minimizes the MSE between the bit and the filter output of user k

$$\text{MSE}_k = E\left[(\boldsymbol{r}^\top \boldsymbol{c}_k - b_k)^2\right] = \boldsymbol{c}_k^\top \boldsymbol{B} \boldsymbol{c}_k - 2\sqrt{p_k}\boldsymbol{c}_k^\top \boldsymbol{s}_k + 1 \tag{10}$$

where $\boldsymbol{B} = \sum_{j=1}^N p_j \boldsymbol{s}_j \boldsymbol{s}_j^\top + \sigma^2 \boldsymbol{I}_L = \boldsymbol{S}\boldsymbol{P}\boldsymbol{S}^\top + \sigma^2 \boldsymbol{I}_L$. The MMSE filter for the kth user is

$$\boldsymbol{c}_k = \arg\min_{\boldsymbol{c}_k} \text{MSE}_k = \sqrt{p_i}\boldsymbol{B}^{-1}\boldsymbol{s}_k \tag{11}$$

Note that the MMSE solution can also be written as

$$\boldsymbol{c}_k = \frac{\sqrt{p_k}\left(\boldsymbol{A}_k + \sigma^2 \boldsymbol{I}_L\right)^{-1} \boldsymbol{s}_k}{1 + p_k \boldsymbol{s}_k^\top \left(\boldsymbol{A}_k + \sigma^2 \boldsymbol{I}_L\right)^{-1} \boldsymbol{s}_k} \tag{12}$$

where $\boldsymbol{A}_k = \sum_{j \neq k} p_j \boldsymbol{s}_j \boldsymbol{s}_j^\top$.

It was found in [3] that the network capacity of a CDMA system with MMSE receivers satisfies

$$N < L\left(1 + \frac{1}{\beta}\right) \tag{13}$$

It was shown in [3] that the network capacity with MMSE receivers is maximized if the signature sequence set is chosen to satisfy (8) if $N \leq L$ and (9) if $N > L$, and if the received powers of the users are chosen to be the same.

It was also shown in [3] that when the matched filters are used as the receivers, the network capacity is again given by (13) and the maximum is achieved by equal powers and signature sequence sets satisfying (8) and (9) for $N \leq L$ and $N > L$, respectively. As was observed in [3], this result is expected because once the signature set is chosen to satisfy (8) and (9), for $N \leq L$ and $N > L$, respectively, the MMSE receivers are scaled matched filters.

Although it is known that the sequences which satisfy (8) for $N \leq L$ and (9) for $N > L$ maximize the capacity of a single cell CDMA system, no simple algorithmic scheme is known to construct these sequence sets for general N and L. Note that when $N \leq L$, a simple Gram-Schmidt orthogonalization procedure would yield N orthogonal vectors starting with N linearly independent vectors. For $N > L$, no such simple construction scheme exists.

2 WBE Sequences and a Simple Observation

Welch developed lower bounds for the $2k$th power of the maximum correlation among a set of N unit energy sequences [6]. These lower bounds on the $2k$th powers of the maximum correlation were actually obtained from a lower bound on the sum of the $2k$th powers of the correlations of a sequence set.

$$\sum_{i=1}^{N}\sum_{j=1}^{N} (\boldsymbol{s}_i^\top \boldsymbol{s}_j)^{2k} \geq \frac{N^2}{\binom{L+k-1}{k}} \tag{14}$$

When $k = 1$, Welch's bound reduces to

$$\sum_{i=1}^{N}\sum_{j=1}^{N} (\boldsymbol{s}_i^\top \boldsymbol{s}_j)^2 \geq \frac{N^2}{L} \tag{15}$$

For a simpler derivation of the bound (15), see [7,8]. Note that sequence sets satisfying (9) satisfy the bound in (15) with equality.

Consider a set of vectors, s_i, $i = 1, \cdots, N$ used as signature sequences by N users. Assume that the signals of the users at the base station are received by matched filters and $p_i = p$, for all i. In this case, the mean squared error for the ith user is

$$\mathrm{MSE}_i = E\left[\left(\boldsymbol{r}^\top \boldsymbol{s}_i - b_i\right)^2\right] = \boldsymbol{s}_i^\top \boldsymbol{B}\boldsymbol{s}_i - 2\sqrt{p}\boldsymbol{s}_i^\top \boldsymbol{s}_i + 1 \tag{16}$$

where $\boldsymbol{r}$ is the received signal given in (2) and $\boldsymbol{B} = p\sum_{j=1}^{N} \boldsymbol{s}_j\boldsymbol{s}_j^\top + \sigma^2 \boldsymbol{I}_L$. The total mean squared error in the system is

$$\begin{aligned}\mathrm{MSE} = \sum_{i=1}^{N} \mathrm{MSE}_i &= \sum_{i=1}^{N} \boldsymbol{s}_i^\top \left(p\sum_{j=1}^{N} \boldsymbol{s}_j\boldsymbol{s}_j^\top + \sigma^2 \boldsymbol{I}_L\right) \boldsymbol{s}_i - 2\sqrt{p}\sum_{i=1}^{N} \boldsymbol{s}_i^\top \boldsymbol{s}_i + N \\ &= p\sum_{i=1}^{N}\sum_{j=1}^{N}(\boldsymbol{s}_i^\top \boldsymbol{s}_j)^2 - (2\sqrt{p} - \sigma^2)\sum_{i} \boldsymbol{s}_i^\top \boldsymbol{s}_i + N \end{aligned}\tag{17}$$

Defining total squared correlation (TSC) as

$$\mathrm{TSC} = \sum_{i=1}^{N}\sum_{j=1}^{N}(\boldsymbol{s}_i^\top \boldsymbol{s}_j)^2 \tag{18}$$

we observe that

$$\mathrm{MSE} = p\mathrm{TSC} - (2\sqrt{p} - \sigma^2)\sum_{i=1}^{N} \boldsymbol{s}_i^\top \boldsymbol{s}_i + N \tag{19}$$

Since the signature sequences are restricted to be of unit energy,

$$\mathrm{MSE} = p\mathrm{TSC} + (1 + \sigma^2 - 2\sqrt{p})N \tag{20}$$

and minimizing TSC subject to $\boldsymbol{s}_i^\top \boldsymbol{s}_i = 1$ is equivalent to minimizing MSE subject to the same condition. Since the sequence sets satisfying (8) for $N \leq L$, and (9) for $N > L$ minimize the TSC, they minimize the MSE as well. In other words, orthogonal sequences for $N \leq L$ and WBE sequences for $N > L$ are the *global optimal* solutions of the following two equivalent problems

$$\begin{array}{llll} \min & \mathrm{TSC} & \qquad \min & \mathrm{MSE} \\ \text{s. t.} & \boldsymbol{s}_i^\top \boldsymbol{s}_i = 1 \quad i = 1, \ldots, N & \qquad \text{s. t.} & \boldsymbol{s}_i^\top \boldsymbol{s}_i = 1 \quad i = 1, \ldots, N \end{array} \tag{21}$$

In the following we will show that if we replace any one of the N sequences in a set with the normalized MMSE filter corresponding to that sequence, the resulting set of sequences will have a smaller TSC. This simple observation will constitute the basis for the iterative algorithm we will propose in the next section.

We first separate the terms that depend on a particular signature sequence, $\boldsymbol{s}_k$, in the TSC expression

$$\mathrm{TSC} = (\boldsymbol{s}_k^\top \boldsymbol{s}_k)^2 + 2\boldsymbol{s}_k^\top \left(\sum_{j\neq k} \boldsymbol{s}_j\boldsymbol{s}_j^\top\right) \boldsymbol{s}_k + \sum_{i\neq k}\sum_{j\neq k}(\boldsymbol{s}_i^\top \boldsymbol{s}_j)^2 \tag{22}$$

Since we will always restrict ourselves to unit energy signature sequences, i.e., $\boldsymbol{s}_k^\top \boldsymbol{s}_k = 1$, we can add $2a^2 s_k^\top s_k$ and subtract $2a^2$ in the above expression to get

$$\mathrm{TSC} = (s_k^\top s_k)^2 + 2 s_k^\top \left(\sum_{j \neq k} s_j s_j^\top + a^2 I_L \right) s_k + \sum_{i \neq k} \sum_{j \neq k} (s_i^\top s_j)^2 - 2a^2 \tag{23}$$

Let us replace the signature sequence of user k with the unit energy vector

$$c_k = \frac{\left(A_k + a^2 I_L\right)^{-1} s_k}{\left[s_k^\top \left(A_k + a^2 I_L\right)^{-2} s_k \right]^{1/2}} \tag{24}$$

where $A_k = \sum_{j \neq k} s_j s_j^\top$. Thus we map the set of signature sequences S to a new set of signatures

$$\bar{S} = [s_1, s_2, \cdots, s_{k-1}, c_k, s_{k+1}, \cdots, s_N] \tag{25}$$

Note that c_k is the normalized MMSE filter for user k. Also note that this MMSE filter is a generalized one; c_k is the normalized MMSE filter for user k in a CDMA system where all other users transmit with signature sequences s_j where $j \neq k$, all users have received powers $p_i = 1$ and the variance of the AWGN is a^2. It will be apparent in what follows that any generalized normalized MMSE filter will be as good as any other in terms of constructing optimum sequences.

The total squared correlation of set S is

$$\mathrm{TSC} = (s_k^\top s_k)^2 + 2 s_k^\top \left(A_k + a^2 I_L \right) s_k + C - 2a^2 \tag{26}$$

where $C = \sum_{i \neq k} \sum_{j \neq k} (s_i^\top s_j)^2$ represents the squared correlation terms that are not affected by the update: $s_k \rightarrow c_k$. Equations (23) and (24) imply that the modified signature set $\bar{S}$ has total squared correlation

$$\overline{\mathrm{TSC}} = (c_k^\top c_k)^2 + 2 \frac{s_k^\top \left(A_k + a^2 I_L\right)^{-1} s_k}{s_k^\top \left(A_k + a^2 I_L\right)^{-2} s_k} + C - 2a^2 \tag{27}$$

Note that by the nature of the mapping in (24), $c_k^\top c_k = s_k^\top s_k = 1$. The following theorem verifies that replacing a particular sequence with its normalized MMSE receiver cannot increase the total squared correlation of the set. In other words, in terms of the TSC (or MSE) criteria, $\bar{S}$ is a *better* set of signature sequences than S.

Theorem 1 *If any signature s_k is replaced with the normalized MMSE receiver for that vector, c_k, the total squared correlation decreases: $\overline{TSC} \leq TSC$. The inequality is satisfied with equality, i.e., $\overline{TSC} = TSC$, iff $c_k = s_k$.*

Due to space limitations, the proof of Theorem 1 and all other proofs will be omitted here. They can be found in [9].

3 An Iterative Algorithm

We observed in the previous section that given a set of unit energy vectors, if any one of these vectors is replaced with the corresponding normalized MMSE vector, then the TSC of the set decreases. We start with N unit length vectors $S(0) = [s_1(0), \cdots, s_N(0)]$ at time 0. At iteration $(n+1)$ the algorithm replaces the vectors $S(n) = [s_1(n), \cdots, s_N(n)]$ with their corresponding normalized MMSE filters one by one, and yields $S(n+1) = [s_1(n+1), \cdots, s_N(n+1)]$. A complete iteration includes N intermediate steps. At the

kth intermediate step in iteration $(n+1)$, the first $(k-1)$ vectors have already been updated and the current vector set is

$$S_{k-1}(n+1) = [s_1(n+1), \cdots, s_{k-1}(n+1), s_k(n), s_{k+1}(n), \cdots, s_N(n)] \tag{28}$$

The kth vector is then updated according to

$$s_k(n+1) = \frac{\left(A_k(n+1) + a^2 I_L\right)^{-1} s_k(n)}{\left[s_k^\top(n) \left(A_k(n+1) + a^2 I_L\right)^{-2} s_k(n)\right]^{1/2}} \tag{29}$$

to yield the vector set

$$S_k(n+1) = [s_1(n+1), \cdots s_{k-1}(n+1), s_k(n+1), s_{k+1}(n), \cdots, s_N(n)] \tag{30}$$

The matrix $A_k(n+1)$ in (29) is given as

$$A_k(n+1) = \sum_{j<k} s_j(n+1) s_j^\top(n+1) + \sum_{j>k} s_j(n) s_j^\top(n) \tag{31}$$

4 Convergence of the Proposed Algorithm

Let $\mathrm{TSC}_k(n)$ denote the TSC of the set $S_k(n)$ after the kth intermediate step in iteration n. In addition, let $\mathrm{TSC}(n)$ denote the TSC at the end of iteration n for set $S(n)$. As a consequence of Theorem 1 we have

$$\mathrm{TSC}_{k+1}(n) \le \mathrm{TSC}_k(n) \qquad \text{for } k = 0, \cdots, N-1; \quad n \ge 0 \tag{32}$$

By recursive application of (32) we have

$$\mathrm{TSC}(n+1) \le \mathrm{TSC}(n) \tag{33}$$

Note that for $N \le L$, TSC is lower bounded by N, and in general (including the case $N > L$) TSC is lower bounded by N^2/L from Welch's bound. Therefore the monotonically decreasing sequence $\mathrm{TSC}(n)$ converges to a finite number. For $N \le L$, if $\mathrm{TSC}(n)$ converges to N, then $S(n)$ converges to a set of N orthogonal sequences, and for $N > L$ if $\mathrm{TSC}(n)$ converges to N^2/L then $S(n)$ converges to a set of WBE sequences. Note that

$$\mathrm{TSC}(n+1) = \mathrm{TSC}_N(n+1) \le \mathrm{TSC}_{N-1}(n+1) \le \cdots \le \mathrm{TSC}_1(n+1) \le \mathrm{TSC}(n) \tag{34}$$

Thus, at the fixed point where $\mathrm{TSC}(n+1) = \mathrm{TSC}(n)$ we should have

$$\mathrm{TSC}(n+1) = \mathrm{TSC}_N(n+1) = \mathrm{TSC}_{N-1}(n+1) = \cdots = \mathrm{TSC}_1(n+1) = \mathrm{TSC}(n) \tag{35}$$

From Theorem 1, this occurs iff $s_k(n+1) = s_k(n)$ for all k. Thus at the fixed point we must have $S(n+1) = S(n)$.

Let us denote the fixed point set of vectors as $S = [s_1, \cdots, s_N]$. Let $B = SS^\top + a^2 I_L$. From two alternative notations for the MMSE operation in (11) and (12), we will choose to use that given in (11). The fixed point S satisfies

$$B s_i = \phi_i s_i \qquad \text{for } i = 1, \cdots, N \tag{36}$$

In the following two subsections, we will investigate the properties of the fixed point for the cases $N \le L$ and $N > L$ separately. We will identify the conditions under which we can guarantee that the fixed point is actually a set of orthogonal vectors for $N \leq L$ and a set of WBE sequences for $N > L$. We will need the following claim.

Lemma 1 *If the signature s_k is replaced by c_k, at step k of iteration n, then*

(a) $\det(S_{k+1}^\top(n)S_{k+1}(n)) \geq \det(S_k^\top(n)S_k(n))$

(b) $\det(S_{k+1}(n)S_{k+1}^\top(n)) \geq \det(S_k(n)S_k^\top(n))$

Lemma 1(a) implies that if $\det\left(S^\top(0)S(0)\right) > 0$, i.e., $S^\top(0)S(0)$ is invertible, then at the fixed point $\det\left(S^\top S\right) > 0$, i.e., $S^\top S$ is invertible. Similarly, Lemma 1(b) says that if $\det\left(S(0)S^\top(0)\right) > 0$, i.e., $S(0)S^\top(0)$ is invertible, then at the fixed point, $\det\left(SS^\top\right) > 0$, i.e., $SS^\top$ is invertible.

4.1 Properties of the Fixed Point: $N \leq L$

Multiplying the equality in (36) for user i with $s_j^\top$ from left we obtain

$$s_j^\top B s_i = \phi_i s_j^\top s_i \qquad \text{for } i = 1, \cdots, N;\ j = 1, \cdots, N \tag{37}$$

Inserting $B = SS^\top + a^2 I_L$ into (37) and writing it in a matrix form we get

$$S^\top \left(SS^\top + a^2 I_L\right) S = S^\top S \Phi \tag{38}$$

where Φ is an $N \times N$ diagonal matrix with ϕ_i as its ith diagonal element. Equation (38) can be written as

$$\left(S^\top S\right)^2 + a^2 S^\top S = S^\top S \Phi \tag{39}$$

If the algorithm is started with N linearly independent vectors $S(0)$, then by Lemma 1(a), $S^\top S$ will be invertible. Multiplying both sides of (39) with $(S^\top S)^{-1}$ yields

$$S^\top S + a^2 I_N = \Phi \tag{40}$$

Thus, $S^\top S$ is a diagonal matrix. By the nature of the algorithm $s_i^\top s_i = 1$ for all i, and the diagonal elements of $S^\top S$ are all unity. Therefore, at the fixed point for $N \leq L$ we have $S^\top S = I_N$. Note from (40) that at the fixed point all ϕ_i have the same value.

Therefore, for $N \leq L$, if the algorithm is started with N linearly independent signature sequences, the proposed algorithm converges to an orthonormal set of signature sequences.

4.2 Properties of the Fixed point: $N > L$

Multiplying (36) for user i with $s_j^\top$ from left, and multiplying the same equation for user j with $s_i^\top$ from left, we obtain the following two equations

$$s_j^\top B s_i = \phi_i s_j^\top s_i \qquad \text{and} \qquad s_i^\top B s_j = \phi_j s_i^\top s_j \tag{41}$$

Since B is symmetric, we must have

$$\phi_i s_j^\top s_i = \phi_j s_i^\top s_j \tag{42}$$

Thus, if $s_j^\top s_i \neq 0$, then $\phi_i = \phi_j$. If the signature sequence set at the fixed point is not split into two or more orthogonal subsets, then we are guaranteed to have $\phi_i = \phi$ for all

i. Note that some vectors may be orthogonal to each other but as long as we do not have an orthogonal splitting of the vector set into two or more, we can still conclude that all ϕ_i have the same value. As we will see below, we can guarantee that the signature sequence set does not get split into orthogonal subsets at the fixed point by imposing quite mild conditions on the initial set of signature sequences, $\boldsymbol{S}(0)$. Without going into the details of these conditions, let us assume that they are satisfied and $\phi_i = \phi$ for all i. Thus

$$\boldsymbol{s}_j^\top \boldsymbol{B} \boldsymbol{s}_i = \phi \boldsymbol{s}_j^\top \boldsymbol{s}_i \qquad \text{for } i = 1, \cdots, N;\ j = 1, \cdots, N \tag{43}$$

Similar to the case of $N \leq L$, inserting $\boldsymbol{B} = \boldsymbol{S}\boldsymbol{S}^\top + a^2 \boldsymbol{I}_L$ into (43) and writing it in a matrix form we get

$$\boldsymbol{S}^\top \left(\boldsymbol{S}\boldsymbol{S}^\top + a^2 \boldsymbol{I}_L \right) \boldsymbol{S} = \phi \boldsymbol{S}^\top \boldsymbol{S} \tag{44}$$

Multiplying both sides of (44) from left by $\boldsymbol{S}$ and from right by $\boldsymbol{S}^\top$ yields

$$\left(\boldsymbol{S}\boldsymbol{S}^\top \right)^3 + a^2 \left(\boldsymbol{S}\boldsymbol{S}^\top \right)^2 = \phi \left(\boldsymbol{S}\boldsymbol{S}^\top \right)^2 \tag{45}$$

By Lemma 1(b), $\boldsymbol{S}\boldsymbol{S}^\top$ is invertible if the rank of $\boldsymbol{S}(0)$ is L. In this case, multiplying both sides of (45) by $(\boldsymbol{S}\boldsymbol{S}^\top)^{-2}$ yields

$$\boldsymbol{S}\boldsymbol{S}^\top + a^2 \boldsymbol{I}_L = \phi \boldsymbol{I}_L \tag{46}$$

Thus, $\boldsymbol{S}\boldsymbol{S}^\top = \alpha \boldsymbol{I}_L$. Note that $\text{trace}(\boldsymbol{S}\boldsymbol{S}^\top) = \text{trace}(\boldsymbol{S}^\top \boldsymbol{S}) = N$, since $\boldsymbol{s}_i^\top \boldsymbol{s}_i = 1$ for all i. Since $\text{trace}(\alpha \boldsymbol{I}_L) = \alpha L$, we have $\alpha = N/L$ and $\boldsymbol{S}\boldsymbol{S}^\top = (N/L)\boldsymbol{I}_L$.

Recursive application of the following Lemma verifies that for $N > L$, a signature set which is not split into orthogonal subsets at time 0 cannot be split into orthogonal subsets at any time n, including the fixed point.

Lemma 2 *For $N > L$, $\boldsymbol{S}(n)$ has two or more orthogonal subsets iff $\boldsymbol{S}(n-1)$ has two or more orthogonal subsets.*

Therefore, for $N > L$, if the algorithm is started with N signature sequences, L of which are linearly independent, and if the initial set of vectors are chosen such that it does not include two or more orthogonal subsets then the proposed algorithm converges to a set of WBE signature sequences.

5 Optimum Orthogonal Partitioning

We have observed that for $N > L$, our proof of convergence to WBE sequences requires that the initial signatures not be partitioned into orthogonal subsets. When $\boldsymbol{S}(0)$ can be partitioned into orthogonal subsets $\boldsymbol{S}'(0)$ and $\boldsymbol{S}''(0)$, we can view the signal subsets as belonging to orthogonal systems. The MMSE iteration preserves orthogonality among systems. Although TSC in this case still decreases monotonically, the algorithm is not guaranteed to converge to a WBE set and the TSC does not decrease to N^2/L. Note that within each orthogonal system, the signatures converge to either an orthogonal or WBE signature set depending on the number of users and the rank of each set. Denoting the TSC achieved by not partitioning the available bandwidth by TSC_1 and the TSC achieved by bandwidth partitioning by TSC_2, we can show that $\text{TSC}_1 \leq \text{TSC}_2$ [10]. Thus, we can conclude that orthogonal partitioning is suboptimal in general. But under certain conditions $\text{TSC}_1 = \text{TSC}_2$ and an orthogonal splitting can actually be as good as non-splitting. Note that the processing gain L is proportional to the available bandwidth. For a fixed bandwidth W, we can either have a signature sequence set with processing gain

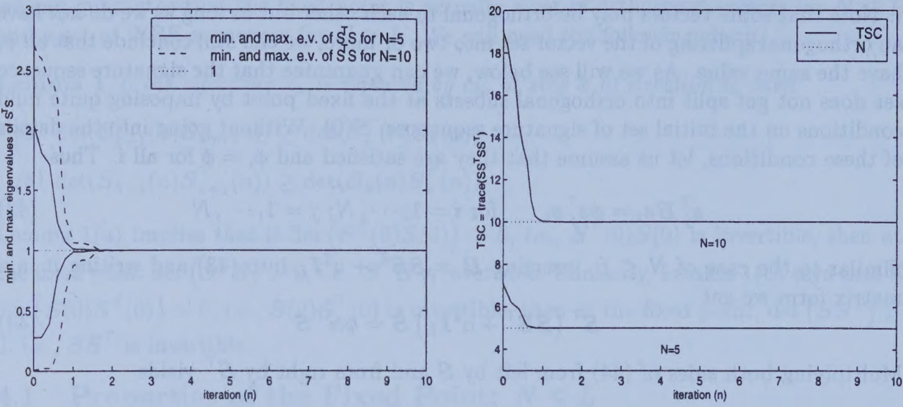


Figure 1: Minimum and maximum eigenvalues of matrix $\boldsymbol{S}^{\top}(n)\boldsymbol{S}(n)$, and TSC$(n)$.

L, or two orthogonal signature sequence sets with processing gains L_1 and $L_2 = L - L_1$ with N_1 and N_2 users, corresponding to bandwidth partitions W_1 and $W_2 = W - W_1$.

It can be shown [9,10] that only in two cases orthogonal partitioning of the signature sequence set yields equally good solutions as not partitioning: if $N \leq L$, any partitioning is as good as the optimal one as long as both spaces have enough dimensionality to assign orthogonal signature sequences to the users; and if $N > L$, we must have the equal loading condition $N_1/L_1 = N/L$.

6 Simulation Results

In this section we present some simple simulation results to verify our analysis. We take the processing gain to be $L = 10$. The initial signature sequences are created randomly. In all the figures, N updates take place between iterations n and $(n+1)$. In each update ith user's signature sequence is replaced with the corresponding normalized MMSE filter, for $i = 1, \cdots, N$. Figure 1 shows the minimum and maximum eigenvalues of the matrix $\boldsymbol{S}^{\top}(n)\boldsymbol{S}(n)$, and TSC$(n)$ as a function of the iteration index for number of users $N = 5, 10$. As expected, the minimum and maximum eigenvalues of $\boldsymbol{S}^{\top}(n)\boldsymbol{S}(n)$ converge to 1 implying that the matrix converges to $\boldsymbol{S}^{\top}\boldsymbol{S} = \boldsymbol{I}_N$, and the TSC converges to N since $N \leq L$.

Figure 2 shows the minimum and maximum eigenvalues of the matrix $\boldsymbol{S}(n)\boldsymbol{S}^{\top}(n)$ and TSC(n) for number of users $N = 20, 30, 40, 50$. As expected, the minimum and maximum eigenvalues of the matrix $\boldsymbol{S}(n)\boldsymbol{S}^{\top}(n)$ converge to N/L implying that the matrix converges to $\boldsymbol{S}\boldsymbol{S}^{\top} = (N/L)\boldsymbol{I}_L$, and the TSC converges to N^2/L, since in these cases $N > L$.

7 Conclusions

Recently it was shown that in order to maximize the capacity of a single cell synchronous CDMA system, if $N \leq L$ then N orthonormal sequences, and if $N > L$ then WBE sequences should be used as signature sequences. In this paper we proposed an algorithm which iteratively updates the signature sequences of the users one at a time and converges to the optimum signature sequence sets. The algorithm replaces one user's signature sequence with the normalized MMSE filter of that user at each update. Since the MMSE receiver can be constructed in a distributed fashion, the proposed algorithm can be

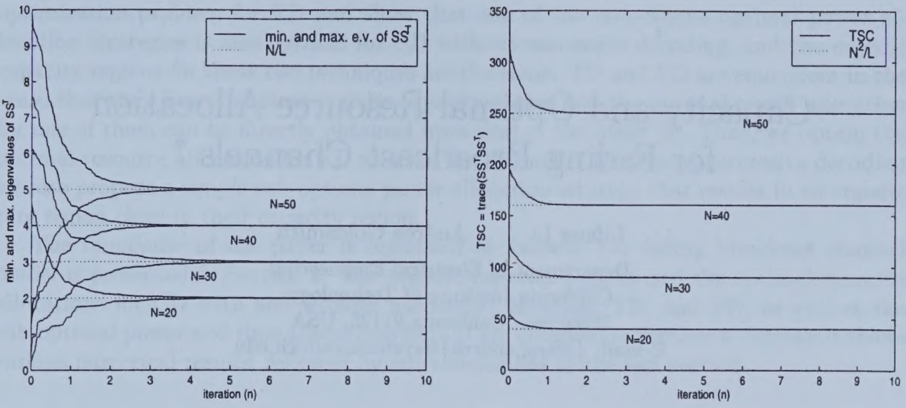


Figure 2: Minimum and maximum eigenvalues of matrix $\boldsymbol{S}^\top(n)\boldsymbol{S}(n)$, and TSC($n$).

implemented distributedly. In this case, each user updates its own signature sequence by making some local measurements and the whole signature sequence set converges to an optimum one.

References

[1] S. Verdú. Capacity region of Gaussian CDMA channels: The symbol-synchronous case. In *Allerton Conf. on Comm., Control and Computing*, pages 1025–1034, October 1986.

[2] M. Rupf and J. L . Massey. Optimum sequence multisets for synchronous code-division multiple-access channels. *IEEE Trans. on Info. Theory*, IT-40(4):1261–1266, July 1994.

[3] P. Viswanath, V. Anatharam, and D. Tse. Optimal sequences, power control and capacity of spread-spectrum systems with multiuser receivers. Preprint. Submitted to IEEE Transactions on Information Theory.

[4] Z. Xie, R. T. Short, and C. K. Rushforth. A family of suboptimum detectors for coherent multiuser communications. *IEEE JSAC*, 8(4):683–690, May 1990.

[5] U. Madhow and M. L. Honig. MMSE interference suppression for direct-sequence spread-spectrum CDMA. *IEEE Trans. on Comm.*, 42(12):3178–3188, December 1994.

[6] L. R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Trans. on Info. Theory*, IT-20(3):397–399, May 1974.

[7] J. L. Massey. On welch's bound for the correlation of a sequence set. In *IEEE International Symposium on Information Theory (ISIT'91)*, page 385, 1991.

[8] J. L . Massey and T. Mittelholzer. Welch's bound and sequence sets for code-division multiple-access systems. *Sequences II: Methods in Communication, Security and Computer Science.* R. Capocelli, A. De Santis and U. Vaccaro, Eds. New York: Springer-Verlag, 1991.

[9] S. Ulukus and R. D. Yates. Interference avoidance in CDMA systems by iterative and distributed signature sequence update. Preprint. To be submitted.

[10] S. Ulukus. Power control, multiuser detection and interference avoidance in CDMA systems. Ph.D. Thesis, Dept. of ECE, Rutgers University, Piscataway, NJ. July 1998.

Capacity and Optimal Resource Allocation for Fading Broadcast Channels *

Lifang Li Andrea Goldsmith
Department of Electrical Engineering
California Institute of Technology
Pasadena, California 91125, USA
E-mail: {lifang,andrea}@systems.caltech.edu

Abstract

We derive the ergodic (Shannon) capacity region of an M-user broadcast fading channel for code-division (CD), time-division (TD) and frequency-division (FD), assuming that both the transmitter and the receivers have perfect channel side information. It is shown that by allowing dynamic resource allocation, TD, FD and CD without successive decoding have the same ergodic capacity region and optimal CD has a larger one. Optimal resource allocation policies are obtained for these different spectrum-sharing techniques. A simple sub-optimal policy is also proposed for TD and CD without successive decoding which results in a rate region quite close to their ergodic capacity region. Numerical results are provided for different fading broadcast channels.

1 Introduction

Wireless communication channels vary over time due to user mobility. Dynamic resource allocation over these channels such as variable power, variable rate and variable bandwidth can significantly improve communication reliability and efficiency. Using optimal dynamic power and rate allocation strategies, the ergodic (Shannon) capacity of a single-user fading channel and a fading multiple access channel (MAC) with channel side information (CSI) at both the transmitter(s) and the receiver are obtained in [1] and [2], respectively[1]. The corresponding optimal adaptive power allocation strategies for these channels can be interpreted as "water-filling" over time (or equivalently, over the fading states).

In this paper, we derive the ergodic capacity of an M-user fading broadcast channel with transmitter and receiver CSI and obtain the corresponding optimal resource allocation strategy for code-division (CD) with and without successive decoding, time-division (TD), and frequency-division (FD). The optimal power allocation which achieves the boundary of the ergodic capacity region is derived by solving an optimization problem over a set of time-invariant additive white Gaussian noise (AWGN) broadcast channels with a total average transmit power constraint. For CD with successive decoding, the optimization problem is similar to that of the parallel Gaussian broadcast channels discussed in [3, 4] and is solved directly by applying the results therein. We solve the

---

*This work was supported by NSF Career Award NCR-9501452 and by a grant from Pacific Bell.

[1] The Shannon capacity of a fading channel is called "throughput capacity" in [2].

optimization problem for TD and show that one of the non-unique optimal power allocation strategies is also optimal for CD without successive decoding, and the ergodic capacity regions for these two techniques are the same. TD and FD are equivalent in the sense that they have the same ergodic capacity region and the optimal power allocation for one of them can be directly obtained from that of the other [5]. Thus, we obtain the optimal resource allocation for FD as well. For TD and CD without successive decoding we also propose a simple sub-optimal power allocation strategy that results in an ergodic rate region close to their capacity region.

The remainder of the paper is organized as follows: the fading broadcast channel model is presented in Section 2. The ergodic capacity regions and the optimal resource allocations for CD with and without successive decoding, TD, and FD, as well as the sub-optimal power and time allocations for TD are obtained in Section 3. Section 4 shows various numerical results, followed by our conclusions in the last section.

2 The Fading Broadcast Channel

We consider a discrete-time M-user broadcast fading channel model where the signal source $X[i]$ is composed of M independent information sources and the broadcast channel consists of M independent fading sub-channels. The time-varying sub-channel gains are denoted as $\sqrt{g_1[i]}, \sqrt{g_2[i]}, \ldots, \sqrt{g_M[i]}$, and the Gaussian noises of these sub-channels are denoted as $z_1[i], z_2[i], \ldots, z_M[i]$. Let $\bar{P}$ be the total average transmit power, B the received signal bandwidth, and $N_j/2$ the noise density of $z_j[i]$, $j = 1, 2, \ldots, M$. Since the time-varying received SNR $\gamma_j[i] = \bar{P}g_j[i]/(N_jB)$, $j = 1, 2, \ldots, M$, if we define[2] $n_j[i] = N_j/g_j[i]$ we have $\gamma_j[i] = \bar{P}/(n_j[i]B)$.

For a slowly time-varying broadcast channel, we assume that $\{n_j[i]\}_{j=1}^{M}$ are known to the transmitter and all the M receivers at time i. Thus, the transmitter can vary the transmit power $P_j[i]$ for each user relative to the noise density vector $\mathbf{n}[i] = (n_1[i], n_2[i], \ldots, n_M[i])$, subject only to the average power constraint $\bar{P}$. For TD or FD, it can also vary the fraction of transmission time or bandwidth $\tau_j[i]$ assigned to each user, subject to the constraint $\sum_{j=1}^{M} \tau_j[i] = 1$ for all i. For CD, the superposition code can be varied at each transmission. Since every receiver knows the noise density vector $\mathbf{n}[i]$, they can decode their individual signals by successive decoding based on the known resource allocation strategy given the M noise densities. In practice, it is necessary to send the transmitter strategy to each receiver through either a header on the transmitted data or a pilot tone. We call $\mathbf{n}[i]$ the joint fading process and denote $\mathcal{N}$ as the set of all possible joint fading states. $F(\mathbf{n})$ denotes a given cumulative distribution function (c.d.f.) on $\mathcal{N}$.

3 Ergodic Capacity Regions

Under the assumption that both the transmitters and the receiver have perfect CSI, the ergodic capacity region of a fading MAC is derived in [2] by exploiting its special polymatroidal structure. In that work the optimal resource allocation scheme is obtained by solving a family of optimization problems over a set of parallel Gaussian MACs, one for each fading state. In this section we derive the ergodic capacity region for the fading broadcast channel under the same assumption about CSI. The corresponding optimal resource allocation strategy is obtained by optimizing over a set of parallel Gaussian broadcast channels for CD with and without successive decoding and for TD. For FD it is shown that the ergodic capacity region is the same as for TD and the corresponding optimal power and bandwidth allocation policy can be derived directly from that of TD [5].

---

[2] Note that when $g_j[i] = 0$ for some j, the broadcast channel can be viewed as consisting of less than M sub-channels.

3.1 CD

We first consider superposition coding and successive decoding where, in each joint fading state, the M-user broadcast channel can be viewed as a degraded Gaussian broadcast channel with noise densities $n_1[i], n_2[i], \ldots, n_M[i]$ and the multiresolution signal constellation is optimized relative to these instantaneous noise densities. Given a power allocation policy $\mathcal{P}$, let $P_j(\mathbf{n})$ be the transmit power allocated to User j for the joint fading state $\mathbf{n} = (n_1, n_2, \ldots, n_M)$ and denote $\mathcal{F}$ as the set of all possible power policies satisfying the average power constraint $E_{\mathbf{n}}\left[\sum_{j=1}^{M} P_j(\mathbf{n})\right] \leq \bar{P}$, where $E[\cdot]$ denotes the expectation function. For simplicity, assume that the stationary distribution of the fading processes have a continuous density, i.e., $Pr\{n_i = n_j\} = 0$, $\forall i \neq j$, then the ergodic capacity region is [6]:

$$\mathcal{C}(\bar{P}) = \bigcup_{\mathcal{P} \in \mathcal{F}} \mathcal{C}_{CD}(\mathcal{P}), \tag{1}$$

where $\mathcal{C}_{CD}(\mathcal{P}) = \left\{R_j \leq E_{\mathbf{n}}\left[B \log\left(1 + \frac{P_j(\mathbf{n})}{n_j B + \sum_{i=1}^{M} P_i(\mathbf{n}) \mathbf{1}[n_j > n_i]}\right)\right], \quad j = 1, 2, \cdots, M\right\}$ and $\mathbf{1}[\cdot]$ denotes the indicator function ($\mathbf{1}[x] = 1$ if x is true and zero otherwise). Since this capacity region is convex [6], $\forall \boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_M) \in \Re_+^M$, if a rate vector $\mathbf{R} = (R_1, R_2, \ldots, R_M)$ is a solution to $\max\limits_{\mathbf{R} \in \mathcal{C}(\bar{P})} \boldsymbol{\mu}\boldsymbol{R}$, it will be on the boundary surface of $\mathcal{C}(\bar{P})$. The maximization problem is equivalent to

$$\begin{cases} \max\limits_{\mathbf{P}(\mathbf{n})} \left\{ E_{\mathbf{n}}\left[\sum_{j=1}^{M} \mu_j \ln\left(1 + \frac{P_j(\mathbf{n})}{n_j B + \sum_{i=1}^{M} P_i(\mathbf{n}) \mathbf{1}[n_j > n_i]}\right)\right] - \lambda E_{\mathbf{n}}\left[\sum_{j=1}^{M} P_j(\mathbf{n})\right] \right\} \\ \text{subject to: } E_{\mathbf{n}}\left[\sum_{j=1}^{M} P_j(\mathbf{n})\right] = \bar{P}, \end{cases} \tag{2}$$

where λ is the Lagrangian multiplier and $\mathbf{P}(\mathbf{n}) = [P_1(\mathbf{n}), P_2(\mathbf{n}), \ldots, P_M(\mathbf{n})]$. In (2), μ_i can be viewed as a weighting parameter proportional to the priority of User i ($i = 1, 2, \ldots, M$). The problem in (2) is quite similar to that of the parallel AWGN broadcast channels discussed in [3, 4] and its solution is obtained by applying the results therein. For each fading state $\mathbf{n}$, let the permutation $\pi(i)$ be defined such that $n_{\pi(i)}$ is the ith smallest in the set $\{n_1, n_2, \ldots, n_M\}$. As described below, the optimal power allocation procedure for state $\mathbf{n}$ is essentially "water-filling" [3]:
Step 1: Do not consider assigning power to those users $\pi(i)$ for which $\frac{\mu_{\pi(i)}}{n_{\pi(i)}} \leq \frac{\mu_{\pi(i+1)}}{n_{\pi(i+1)}}$.
Step 2: Define $P_{\pi(1)} = \min\left\{\left[\frac{\mu_{\pi(1)}}{\lambda} - n_{\pi(1)} B\right]_+, \min\limits_{i \neq 1} \frac{\mu_{\pi(1)} n_{\pi(i)} B - \mu_{\pi(i)} n_{\pi(1)} B}{\mu_{\pi(i)} - \mu_{\pi(1)}}\right\}$, where $[x]_+ \triangleq \max(x, 0)$ and assign power $P_{\pi(1)}$ to User $\pi(1)$. If $P_{\pi(1)} = \left[\frac{\mu_{\pi(1)}}{\lambda} - n_{\pi(1)B}\right]_+$, the total power for the state $\mathbf{n}$ has been allocated. If not, increase the noises $n_{\pi(i)} B$ by $P_{\pi(1)}$ and return to Step 1.

In the above procedure, $s^* = \frac{1}{\lambda}$ is the "water-filling" power level that satisfies the total power constraint in (2). This optimal power allocation strategy has a greedy interpretation [4]. Specifically, for $z > 0$ and a given fading state $\mathbf{n}$, define $u_j(z) = \frac{\mu_j}{n_j B + z} - \lambda$, since $u_j(z) \cdot \delta p$ can be viewed as the marginal increase in the objective function in (2) when a marginal power δp is allocated to User j, the optimal solution is obtained greedily by allocating power δp to the user with the largest positive increase $u_j(z) \cdot \delta p$. If no such user exists, no power is assigned to any user. For example, in a two-user system, in order that $\max\limits_{j\in\{1,2\}} \{u_j(z)\} > 0$, z must be no larger than $\max\limits_{j\in\{1,2\}} \left[\frac{\mu_j}{\lambda} - n_jB\right]_+$. Assuming that $\mu_1 < \mu_2$, if $\frac{\mu_1}{n_1} \leq \frac{\mu_2}{n_2}$, then $\forall z > 0$, $u_1(z) < u_2(z)$, which means no power should be assigned to User 1; if $\frac{\mu_1}{n_1} > \frac{\mu_2}{n_2}$, $u_1(z)$ and $u_2(z)$ will cross each other once at $z_0 \triangleq \frac{\mu_1 n_2 B - \mu_2 n_1 B}{\mu_2 - \mu_1}$. In this case, if $z_0 < z_2 \triangleq \frac{\mu_2}{\lambda} - n_2B$, i.e., $\lambda < \frac{\mu_2-\mu_1}{n_2B-n_1B}$, then for $0 < z < z_0$, $u_1(z) > u_2(z)$ and for $z_0 < z < z_2$, $u_2(z) > u_1(z)$; otherwise $u_1(z) > u_2(z)$ for $0 < z < z_1 \triangleq \frac{\mu_1}{\lambda} - n_1B$. This is why the optimal power policy for each fading state $\mathbf{n}$ is as follows when $\mu_1 < \mu_2$: if $\frac{\mu_1}{n_1} \leq \frac{\mu_2}{n_2}$, then $P_1^*(\mathbf{n}) = 0$, $P_2^*(\mathbf{n}) = \left[\frac{\mu_2}{\lambda} - n_2B\right]_+$; if $\frac{\mu_1}{n_1} > \frac{\mu_2}{n_2}$, then: a) if $\lambda < \frac{\mu_2-\mu_1}{n_2B-n_1B}$, $P_1^*(\mathbf{n}) = \frac{\mu_1 n_2 B - \mu_2 n_1 B}{\mu_2 - \mu_1}$, $P_2^*(\mathbf{n}) = \left[\frac{\mu_2}{\lambda} - n_2B - \frac{\mu_1 n_2 B - \mu_2 n_1 B}{\mu_2 - \mu_1}\right]_+$; b) if $\lambda \geq \frac{\mu_2-\mu_1}{n_2B-n_1B}$, $P_1^*(\mathbf{n}) = \left[\frac{\mu_1}{\lambda} - n_1B\right]_+$, $P_2^*(\mathbf{n}) = 0$. When $\mu_1 > \mu_2$, the power policy for User 1 and User 2 is similarly obtained as appropriate substitutions are made for all subscripts. When $\mu_1 = \mu_2$, the power policy is simplified to the following: if $n_1 \leq n_2$, then $P_1^*(\mathbf{n}) = \left[\frac{\mu_1}{\lambda} - n_1B\right]_+$, $P_2^*(\mathbf{n}) = 0$; otherwise $P_1^*(\mathbf{n}) = 0$, $P_2^*(\mathbf{n}) = \left[\frac{\mu_2}{\lambda} - n_2B\right]_+$.

3.2 TD

Now we consider the time-division case where, in each fading state $\mathbf{n}$, the information for the M users will be sent at their time-slots which are pre-assigned by the transmitter. For a given power and time allocation policy $\mathcal{P}$, let $P_j(\mathbf{n})$ and $\tau_j(\mathbf{n})$ $(0 \leq \tau_j(\mathbf{n}) \leq 1)$ be the transmit power and fraction of transmission time allocated to User j $(j = 1, 2, \ldots, M)$, respectively, and let $\mathcal{F}$ be the set of all such possible power and time allocation policies satisfying $E_{\mathbf{n}}\left[\sum_{j=1}^M \tau_j(\mathbf{n})P_j(\mathbf{n})\right] \leq \bar{P}$, $\sum_{j=1}^M \tau_j(\mathbf{n}) = 1$, $\forall \mathbf{n} \in \mathcal{N}$. Then the capacity region for this variable power and variable transmission time scheme is [6]:

$$\mathcal{C}(\bar{P}) = \bigcup_{\mathcal{P}\in\mathcal{F}} \mathcal{C}_{TD}(\mathcal{P}), \tag{3}$$

where $\mathcal{C}_{TD}(\mathcal{P}) = \left\{R_j \leq E_{\mathbf{n}}\left[\tau_j(\mathbf{n})B\log\left(1 + \frac{P_j(\mathbf{n})}{n_jB}\right)\right], j = 1, 2, \cdots, M\right\}$. Due to the convexity of this capacity region [6], $\forall \boldsymbol{\mu} \in \Re_+^M$, if a rate vector $\mathbf{R}$ is a solution to $\max\limits_{\mathbf{R}\in\mathcal{C}(\bar{P})} \boldsymbol{\mu}\mathbf{R}$, it will be on the boundary surface of $\mathcal{C}(\bar{P})$. The maximization problem is equivalent to $\max\limits_{P(\mathbf{n})} E_{\mathbf{n}}[J(P(\mathbf{n}))] - \lambda E_{\mathbf{n}}[P(\mathbf{n})]$ subject to $E_{\mathbf{n}}[P(\mathbf{n})] = \bar{P}$, where λ is the Lagrangian multiplier and

$$\begin{cases} J(P(\mathbf{n})) \triangleq \max\limits_{\boldsymbol{\tau}(\mathbf{n}),\mathbf{P}(\mathbf{n})} \sum_{i=1}^M \tau_i(\mathbf{n}) f_i(P_i(\mathbf{n})) \\ \text{subject to } \sum_{i=1}^M \tau_i(\mathbf{n}) = 1 \text{ and } \sum_{i=1}^M \tau_i(\mathbf{n})P_i(\mathbf{n}) = P(\mathbf{n}), \end{cases} \tag{4}$$

$f_i(x) \triangleq \mu_i \ln[1 + x/(n_iB)]$, $\boldsymbol{\tau}(\mathbf{n}) = [\tau_1(\mathbf{n}), \tau_2(\mathbf{n}), \cdots, \tau_M(\mathbf{n})]$. Thus, for a given fading state $\mathbf{n}$, the optimal power $P^*(\mathbf{n})$ satisfies $J'(P^*(\mathbf{n})) - \lambda = 0$, where $J'(x)$ denotes the derivative of $J(x)$ with respect to x. $\forall i, j$, $\mu_i < \mu_j$, it can be easily shown that if $\frac{\mu_i}{n_i} \leq \frac{\mu_j}{n_j}$ then $\forall x > 0$, $f_i(x) < f_j(x)$ and the optimal $\tau_i^*(\mathbf{n})$ and $P_i^*(\mathbf{n})$ are $\tau_i^*(\mathbf{n}) = 0$, $P_i^*(\mathbf{n}) = 0$; otherwise $f_i(x)$ and $f_j(x)$ will cross each other once at some positive x [6].

For example, in the two-user case, when $\mu_1 < \mu_2$, we have the following two cases: 1) if $\frac{\mu_1}{n_1} \leq \frac{\mu_2}{n_2}$, then $J(P(\mathbf{n})) = f_2(P(\mathbf{n}))$. Thus, $\tau_1^*(\mathbf{n}) = 0$, $\tau_2^*(\mathbf{n}) = 1$, $P_1^*(\mathbf{n}) = 0$,

$P_2^*(\mathbf{n}) = P^*(\mathbf{n}) = \left[\frac{\mu_2}{\lambda} - n_2 B\right]_+$;
2) if $\frac{\mu_1}{n_1} > \frac{\mu_2}{n_2}$, as shown in Figure 1[3], it is clear that [6]

$$J(P(\mathbf{n})) = \begin{cases} f_1(P(\mathbf{n})) & \text{if } J'(P(\mathbf{n})) > \lambda_0 , \\ f_1(P_a(\mathbf{n})) + \lambda_0(P(\mathbf{n}) - P_a(\mathbf{n})) & \text{if } J'(P(\mathbf{n})) = \lambda_0 , \\ f_2(P(\mathbf{n})) & \text{if } J'(P(\mathbf{n})) < \lambda_0 , \end{cases} \tag{5}$$

where λ_0 is the tangent of the straight line in this figure and it satisfies $h(\lambda_0, \mathbf{n}) = 0$,

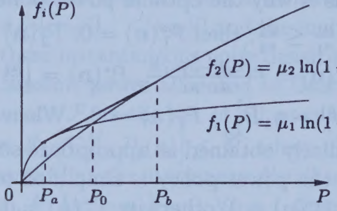


Figure 1: The functions $f_1(P)$ and $f_2(P)$ when $\frac{\mu_1}{n_1} > \frac{\mu_2}{n_2}$.

$h(x, \mathbf{n}) \triangleq (n_2 - n_1)Bx + \left(\mu_2 \ln \frac{\mu_2}{x n_2 B} - \mu_1 \ln \frac{\mu_1}{x n_1 B}\right) - (\mu_2 - \mu_1)$. Therefore, since $J'(P^*(\mathbf{n})) = \lambda$, a) if $\lambda > \lambda_0$, then $\tau_1^*(\mathbf{n}) = 1$, $\tau_2^*(\mathbf{n}) = 0$, $P_1^*(\mathbf{n}) = P^*(\mathbf{n}) = \left[\frac{\mu_1}{\lambda} - n_1 B\right]_+$, $P_2^*(\mathbf{n}) = 0$; b) if $\lambda < \lambda_0$, then $\tau_1^*(\mathbf{n}) = 0$, $\tau_2^*(\mathbf{n}) = 1$, $P_1^*(\mathbf{n}) = 0$, $P_2^*(\mathbf{n}) = \left[\frac{\mu_2}{\lambda} - n_2 B\right]_+$; c) if $\lambda = \lambda_0$, then $\tau_1^*(\mathbf{n}) = \tau_0^*$, $\tau_2^*(\mathbf{n}) = 1 - \tau_0^*$, $P_1^*(\mathbf{n}) = \left[\frac{\mu_1}{\lambda} - n_1 B\right]_+$, $P_2^*(\mathbf{n}) = \left[\frac{\mu_2}{\lambda} - n_2 B\right]_+$, where λ and τ_0^* are determined by the average power constraint $E_{\mathbf{n}}\left[\sum_{i=1}^2 \tau_i^*(\mathbf{n}) P_i^*(\mathbf{n})\right] = \bar{P}$ and they may not be unique. When $\mu_1 > \mu_2$, the optimal power policy for User 1 and User 2 is similarly derived as appropriate substitutions are made for all subscripts. When $\mu_1 = \mu_2$, the optimal power policy is the same as that for CD discussed at the end of Section 3.1.

The optimal power and time allocation policy for the two-user case can be generalized to the M-user case ($M > 2$). Assuming that $\mu_1 \le \mu_2 \le \cdots \le \mu_M$, first we can remove those users i (i.e. let $\tau_i^*(\mathbf{n}) = 0$, $P_i(\mathbf{n})^* = 0$) that satisfy $\frac{\mu_i}{n_i} \le \frac{\mu_{i+1}}{n_{i+1}}$ or satisfy $\mu_{i-1} = \mu_i$ and $\frac{\mu_i}{n_i} < \frac{\mu_{i-1}}{n_{i-1}}$. Next, let $m = 1$ and the following procedure describes the optimal policy for the remaining users in the state $\mathbf{n}$:
Step 1: Denote the number of remaining users as K and define the permutation $\pi(i)$ such that $\mu_{\pi(i)}$ is the ith smallest in the K remaining users, then due to the removal criterion, it must be true that $\frac{\mu_{\pi(1)}}{n_{\pi(1)}} > \frac{\mu_{\pi(2)}}{n_{\pi(2)}} > \cdots > \frac{\mu_{\pi(K)}}{n_{\pi(K)}}$.
Step 2: Let $\pi_0(m) = \pi(1)$. If $K < 2$, go to Step 4; otherwise, go to Step 3.
Step 3: For $1 \le i \le K - 1$, define $h_i(x, \mathbf{n}) \triangleq \left[\mu_{\pi(i+1)} \ln \frac{\mu_{\pi(i+1)}}{x n_{\pi(i+1)} B} - \mu_{\pi(1)} \ln \frac{\mu_{\pi(1)}}{x n_{\pi(1)} B}\right] + (n_{\pi(i+1)} - n_{\pi(1)})Bx - (\mu_{\pi(i+1)} - \mu_{\pi(1)})$, and let x_i be the solution to $h_i(x, \mathbf{n}) = 0$. Let $\lambda_m = max_i\{x_i\}$, $i^* = arg\ max_i\{x_i\}$. Remove those users $\pi(j)$ for which $1 < j \le i^*$ and let $\tau_{\pi(j)}^*(\mathbf{n}) = 0$, $P_{\pi(j)}^*(\mathbf{n}) = 0$. Also remove User $\pi(1)$. Increase m by 1 and return to Step 1.
Step 4: Given $\lambda > 0$, if $m = 1$, then $\tau_{\pi_0(1)}^*(\mathbf{n}) = 1$ and $P_{\pi_0(1)}^*(\mathbf{n}) = \left[\frac{\mu_{\pi_0(1)}}{\lambda} - n_{\pi_0(1)} B\right]_+$; if $m > 1$, by denoting $\lambda_0 \triangleq \infty$ and $\lambda_m \triangleq 0$, we know that there exists a $j \in \{1, 2, \ldots, m\}$

---

[3] In this figure, P, P_a, P_0 and P_b are all functions of $\mathbf{n}$, which is not shown for simplicity.

such that $\lambda_j < \lambda \leq \lambda_{j-1}$, since $0 = \lambda_m < \lambda_{m-1} < \cdots < \lambda_1 < \lambda_0 = \infty$, which results from Step 3. If $\lambda_j < \lambda < \lambda_{j-1}$, then $\tau^*_{\pi_0(j)}(\mathbf{n}) = 1$ and $P^*_{\pi_0(j)}(\mathbf{n}) = \left[\frac{\mu_{\pi_0(j)}}{\lambda} - n_{\pi_0(j)}B\right]_+$; if $\lambda = \lambda_j$, then $\tau^*_{\pi_0(j)}(\mathbf{n}) = \tau^*_0$, $\tau^*_{\pi_0(j+1)}(\mathbf{n}) = 1 - \tau^*_0$, $P^*_{\pi_0(j)}(\mathbf{n}) = \left[\frac{\mu_{\pi_0(j)}}{\lambda} - n_{\pi_0(j)}B\right]_+$ and $P^*_{\pi_0(j+1)}(\mathbf{n}) = \left[\frac{\mu_{\pi_0(j+1)}}{\lambda} - n_{\pi_0(j+1)}B\right]_+$.

In the above procedure, λ and τ^*_0 satisfy the average power constraint

$$E_{\mathbf{n}}\left[\sum_{i=1}^{M} \tau^*_i(\mathbf{n})P^*_i(\mathbf{n})\right] = \bar{P}. \tag{6}$$

This optimal power and time allocation procedure implies that in each fading state $\mathbf{n}$, the information of at most two users is transmitted via time-sharing the channel. If $F(\mathbf{n})$ is continuous, $\forall\lambda > 0$, the probability measure of the set $L_i \triangleq \{\mathbf{n} : h_i(\lambda, \mathbf{n}) = 0\}$ $(i = 1, 2, \cdots, m)$ is zero and λ is uniquely determined by (6), since the value of τ^*_0 is only effective on a set of $\mathbf{n}$ with zero-probability measure. Moreover, in this case, with probability 1, the information of at most a single user is transmitted in each fading state. We therefore propose the following sub-optimal method to choose the single user.

In each state $\mathbf{n}$, $\forall i = 1, 2, \ldots, M$, define $P_i(\mathbf{n}) = \left[\frac{\mu_j}{\lambda} - n_jB\right]_+$, $g_i(x) = \mu_i \ln\left(1 + \frac{x}{n_iB}\right)$. Let $G(\mathbf{n}) = max_i\ g_i(P_i(\mathbf{n}))$, $S = \{i : g_i(P_i(\mathbf{n})) = G(\mathbf{n})\}$, and $i^* \triangleq arg\ min_{i\in S}P_i(\mathbf{n})$. Then, in state $\mathbf{n}$, only the information for User i^* is sent with transmit power $P_{i^*}(\mathbf{n})$. Compared to the optimal power and time allocation policy, the advantage of this sub-optimal scheme is that it is much easier to compute the "water-filling" power level $s^* \triangleq \frac{1}{\lambda}$ using (6). As will be shown in Section 4, the resulting rate region comes very close to that of the optimal policy.

3.3 CD without Successive Decoding

If we consider CD without successive decoding, since each receiver simply treats the signals for other users as interfering noises, the ergodic capacity region is [6]:

$$\mathcal{C}(\bar{P}) = \bigcup_{\mathcal{P}\in\mathcal{F}} \mathcal{C}_{CDWO}(\mathcal{P}), \tag{7}$$

where, as in Section 3.1, for a given power allocation policy $\mathcal{P}$, the transmit power allocated to User j is $P_j(\mathbf{n})$, $j = 1, 2, \ldots, M$, $\mathcal{F}$ denotes the set of all possible power policies satisfying the average power constraint $E_{\mathbf{n}}\left[\sum_{j=1}^{M} P_j(\mathbf{n})\right] \leq \bar{P}$, and $\mathcal{C}_{CDWO}(\mathcal{P}) = \left\{R_j \leq E_{\mathbf{n}}\left[B\log\left(1 + \frac{P_j(\mathbf{n})}{n_jB + \sum_{i=1,i\neq j}^{M} P_i(\mathbf{n})}\right)\right],\ j = 1, 2, \cdots, M\right\}$. To show that $\mathcal{C}(\bar{P})$ in (7) cannot be larger than the capacity region of TD in (3), we give the following lemma, which is proved in [6].

Lemma 3.1 $\forall P_i \geq 0$, $\mu_i \geq 0$, $n_iB > 0$, $i = 1, 2, \ldots, M$,

$$\sum_{i=1}^{M} \mu_i \log\left(1 + \frac{P_i}{n_iB + \sum_{j=1,j\neq i}^{M} P_j}\right) \leq \max_{i\in\{1,2,\ldots,M\}} \left[\mu_i \log\left(1 + \frac{\sum_{j=1}^{M} P_j}{n_iB}\right)\right]. \tag{8}$$

According to this lemma, for a given state $\mathbf{n}$, by denoting $P = \sum_{i=1}^{M} P_i$, $C_i = B\log\left[1 + \frac{P}{n_iB}\right]$, $R_i = B\log\left(1 + \frac{P_i}{n_iB + \sum_{j=1,j\neq i}^{M} P_j}\right)$, and $\mu_i = \frac{B}{C_i}$, $i = 1, 2, \cdots, M$, we have $\sum_{i=1}^{M} R_i/C_i \leq 1$.

Now consider the equal power TD strategy which assigns the power P to each user i for a fraction τ_i of the total transmission time in this state. By denoting $R_i' = \tau_i B \log[1 + P/(n_i B)]$, we have $\sum_{i=1}^{M} R_i'/C_i = 1$, because $\sum_{i=1}^{M} \tau_i = 1$. Therefore, given a fading state $\mathbf{n}$, $\forall P > 0$, the capacity region of the equivalent AWGN broadcast channel for CD without successive decoding is within that for equal power TD and is therefore within that for optimal TD. Consequently, the capacity region of CD without successive decoding in (7) cannot be larger than the capacity region of TD in (3). In fact, the two capacity regions are the same. This is because if $F(\mathbf{n})$ is continuous, the optimal policy for TD implies that with probability 1, no more than one user is using the broadcast channel in each fading state $\mathbf{n}$. Thus, this optimal strategy can be used as a power allocation policy for CD without successive decoding, which results in exactly the same capacity region. Even if $F(\mathbf{n})$ is discontinuous, the two capacity regions are still the same since the capacity region for TD can be achieved by letting τ_0 equal 0 or 1 in its optimal resource allocation policy, which is again a valid power policy for CD without successive decoding. Similarly, the sub-optimal TD scheme proposed at the end of Section 3.2 is also applicable for CD without successive decoding.

4 Numerical Results

In this section, we present numerical results for the two-user ergodic capacity regions of the Rician and Rayleigh fading channels under different spectrum-sharing techniques. In the figures, as discussed in [5], the equal power TD scheme refers to the strategy which assigns the constant transmission power $\bar{P}$ and bandwidth B to User 1 for a fraction τ of the total transmission time, and then to User 2 for the remainder of the transmission. The optimal TD scheme for both the AWGN channel and the fading channels is obtained by allocating different power to the two users. We refer to CD without successive decoding as CDWO. Since TD and FD are equivalent in the sense that they have the same capacity region, all results for TD in the figures also apply for FD.

In Figure 2, the ergodic capacity regions of the Rician and Rayleigh fading broadcast channels are compared to that of the Gaussian broadcast channel using CD, TD, equal power TD, and CDWO techniques. The SNR difference between the two users is 3 dB ($\bar{n}_1 B$ and $\bar{n}_2 B$ denote the average noise variances of User 1's channel and User 2's channel, respectively). The total transmission power is $\bar{P} = 10$ dB and the signal bandwidth $B = 100$ KHz. The ratio of the direct-path power to the scattered-path power in the Rician fading sub-channels is $k = 6$ dB.

In this figure we see that while the single-user ergodic rate in fading is smaller than the rate on the AWGN channel, the two-user capacity regions of both the Rician fading and the Rayleigh fading broadcast channels using either optimal CD or sub-optimal TD techniques in some places dominate that of the AWGN broadcast channel using optimal CD. That is, for the fading broadcast channel, ergodic rate pairs beyond the capacity region of the non-fading broadcast channel can be achieved by applying optimal resource allocation over the joint fading channel states. For simplicity, we calculate the rate region of fading broadcast channels for TD by applying the simple sub-optimal TD power allocation policy. The resulting rate region turns out to be very close to the capacity region for optimal CD. Therefore, the capacity region using the optimal TD power policy will also come very close to that of optimal CD. This observation implies that, due to the small SNR difference between the two users, superposition encoding with successive decoding is not necessary, since time-sharing is near-optimal. For the

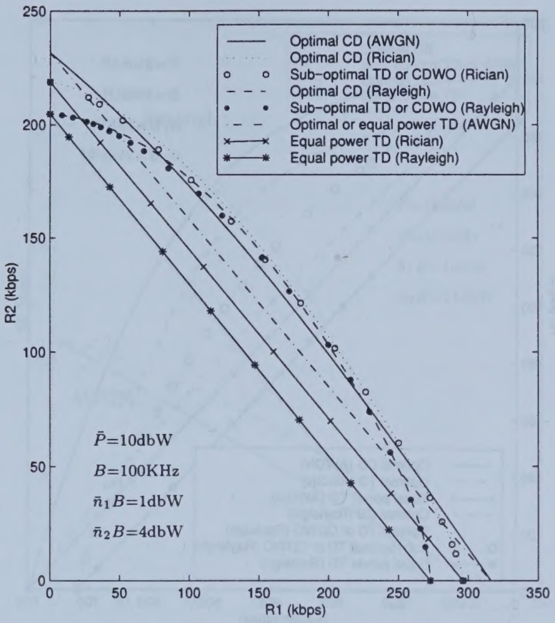


Figure 2: Two-user ergodic capacity region comparison: 3 dB SNR difference.

AWGN broadcast channel, the capacity region boundary of the optimal TD scheme is indistinguishable from the equal power TD straight line, which means that when the two users have a similar channel noise density, constant power allocation is good enough for TD; the CDWO capacity region boundary (omitted from Figure 2 but shown in figures in [5]) includes the two end points of the equal power TD line but is below the straight line due to its convexity. However, for the fading channels, optimal or sub-optimal TD has a much larger capacity region than equal power TD and the capacity region for CDWO is the same as that for TD.

We show in Figure 3 that when the SNR difference between the two users is 20 dB and the total average power is 25 dB, the ergodic capacity region for CD of the Rayleigh fading channel is now completely within that of optimal CD for the AWGN channel. However, optimal TD in fading can achieve some rate pairs far beyond the capacity region of the AWGN broadcast channel using optimal TD, and the sub-optimal TD power policy for Rayleigh fading results in a rate region almost as large as the capacity region with the optimal TD power policy. For both AWGN and fading channels, due to the large SNR difference between the two users, the capacity region for optimal CD is noticeably larger than that for optimal TD, and the capacity region for optimal TD is noticeably larger than that for equal power TD.

Figure 4 shows the case where the SNR difference between the two users is 20 dB and the total average power is only 10 dB. Unlike the previous cases, we see here that the ergodic single-user capacity of User 2 for the Rayleigh fading channel is larger than that for the AWGN channel due to its very low average SNR. Thus, the optimal CD or sub-optimal TD scheme for fading yields a large rate region that is not achievable for the

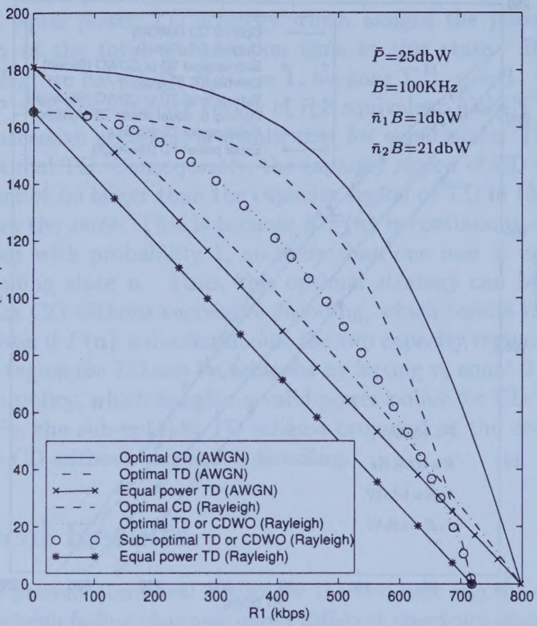


Figure 3: Two-user throughput capacity region comparison: 20dB SNR difference with strong signal power.

AWGN channel. However, as in Figure 3, due to the great SNR difference between the two users, optimal CD results in a capacity region much larger than that for sub-optimal TD and the capacity region for optimal TD is significantly larger than that for equal power TD. This observation holds for both fading and AWGN channels.

5 Conclusions

We have obtained the ergodic capacity region and the optimal dynamic resource allocation strategy for fading broadcast channels with perfect CSI at both the transmitter and the receiver. These results are obtained for CD with and without successive decoding, TD and FD. Comparisons of the capacity regions show that CD with successive decoding has the largest capacity region, while TD and FD are equivalent and they have the same capacity region as CD without successive decoding. For CD without successive decoding, the optimal power policy is to transmit the information of at most one user in each joint fading state. This policy is also optimal for TD, though other strategies which allow at most two users to time-share the channel may also be optimal. When the average channel fading condition for each user is similar, the capacity regions for optimal CD and TD are quite close to each other, However, when each user has an average channel condition quite different from that of the others, using optimal CD can achieve a much larger ergodic capacity region than the other techniques.

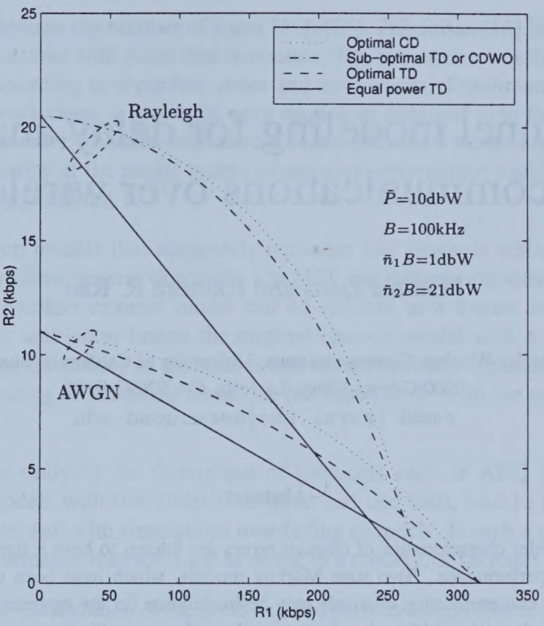


Figure 4: Two-user throughput capacity region comparison: 20dB SNR difference with weak signal power.

References

[1] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inform. Theory*, vol. IT-43, pp. 1986–1992, Nov. 1997.

[2] D. Tse and S. V. Hanly, "Multiple-access fading channels: Part I: Polymatroidal structure, optimal resource allocation and throughput capacities." To appear in IEEE Trans. Inform. Theory.

[3] D. Hughes-Hartogs, "The capacity of a degraded spectral gaussian broadcast channel." Ph. D. Thesis and Technical Report No. 7002-2, Information Systems Laboratory, Center for Systems Research, Stanford University, Jul. 1975.

[4] D. Tse, "Optimal power allocation over parallel gaussian broadcast channels," in *Proc. of International Symp. Inform. Theory, Ulm, Germany*, p. 7, Jun. 1997.

[5] A. J. Goldsmith, "The capacity of downlink fading channels with variable-rate and power," *IEEE Trans. Vehic. Technol*, vol. VT-46, pp. 569–580, Aug. 1997.

[6] L. Li and A. J. Goldsmith, "Capacities of broadcast fading channels with channel side information under different spectrum-sharing techniques: Part I: Ergodic capacities." In Preparation.

On channel modeling for delay analysis of packet communications over wireless links

Michele Zorzi and Ramesh R. Rao

Center for Wireless Communications, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0407
e-mail: `{zorzi,rao}@ece.ucsd.edu`

Abstract

Higher order characteristics of channel errors are known to have a significant impact on protocol performance. Two state Markov models, which have been widely used to model bursty communication channels may be inadequate for the representation of some time-varying channels. Although a larger number of states will almost surely provide a better representation of the channel, the complexity of the model makes the analysis of higher layer protocols difficult. Furthermore, the quality of a particular channel model has much to do with the nature of the metric being computed.

Although a two state Markov channel model leads to throughput predictions of protocols such as ARQ that agree quite well with simulations over Rayleigh fading channels, such a model is found to be inadequate for predicting delay in an ARQ/queueing system that operates over a fading channel. On the other hand, good matches with simulation results are obtained by modeling the error process by means of a Markov model with one good state and two bad states. The model with two bad states results in a richer, non geometric fade duration and can consequently better model the effect of fading from the perspective of delay. The effect of model parameter values on the goodness of the approximation is also discussed.

1 Introduction

For some time now, we have been studying the effect of wireless channel errors on the performance of protocols [1]–[6]. These studies have shown that higher order characteristics of channel errors have a significant impact on performance and that correlations in channel errors can be exploited to reduce energy consumption. These findings suggest that proper modeling of the channel errors is key in being able to accurately predict the performance of a protocol, and motivated us to reexamine the channel models we have used in the past. Markov models have been widely used to model bursty communication channels [7], from the early work of Gilbert [8] to more recent papers studying their applicability to fading channels [1], [9]–[15].

Although more general than the iid (memoryless) error model, the 2-state channel model [1] may be inadequate for the representation of some time-varying channels. One way to overcome this problem is to enlarge the number of states [11]–[15]. Fritchman [16] investigated a finite-state Markov chain model with more than two states. The state space was then partitioned into two groups, corresponding to error-free states and error states. Fritchman's model has been applied by many researchers to represent error sequences obtained over the fading channels. Although a large number of states will almost surely provide a better representation of the channel, the complexity of the model makes subsequent performance analysis of higher layer protocols increasingly difficult.

The quest to develop models that adequately represent real channels while being mathematically tractable has taken several directions. In [17], the authors explored conditions under which an N-state Markov channel model can be reduced to a 2-state model and ways to stochastically upper and lower bound the original channel model with a 2-state Markovian model. In comparing the quality of the bounds that result it became apparent that the goodness of a particular bounding or modeling technique has much to do with the nature of the metric being computed.

In particular, while studying the throughput of protocols such as ARQ [2] and TCP [18], Markov channel models with two states (one good and one bad), lead to throughput predictions that agree quite well with simulations over fading channels. Is such a model adequate for studying other performance metrics, such as delay, in a model that includes queues?

In a two state model, the duration of the good and bad periods are geometrically distributed. As a result, the distribution of the residual life (time to end of duration) of both bad and good periods will be memoryless and have the same distribution as their respective typical distributions. In fact, iid models also exhibit this property with the additional constraint that the parameters of the two geometric distributions are complements of each other. The difference between an iid model and a two state model is that in the latter case, the two geometric distributions can be parameterized by arbitrary parameters. One may anticipate that realistic channels are unlikely to produce bad and good periods that are memoryless and therefore one is forced to consider higher order chains.

In this paper we study a simple example that reinforces these observations. We consider a simple ARQ/queueing model, operating over a Rayleigh fading channel (as in the classic model described in [19, 20]) and show that good matches with simulation results can be obtained by modeling the error process by means of a Markov model with three states (one good and two bad). The additional flexibility introduced by the third state allows us to adequately predict the complementary distribution of the packet delay for the cases considered.

What appears to be happening over the fading channel is that packets awaiting transmission when the channel is in a deep fade see atypical durations of the residual life of the fade. The model with two bad states results in a richer, non geometric fade duration and can consequently better model the effect of fading from the perspective of delay. The effect of model parameter values on the goodness of the approximation is also discussed.

2 System Model

We assume a wireless link, subject to bursty errors, with a relatively high data rate, of the order of 1 Mbps. The time axis is slotted, and transmissions occur in blocks (packets). We focus on a single pair of communicating users. In each slot, packets are independently generated at the transmitter according to a Bernoulli arrival process with rate λ, and queued in a buffer awaiting transmission.

A simple ARQ scheme is used to counteract the effect of packet errors. A perfect and instantaneous feedback channel is assumed, so that at the end of each slot the transmitter knows exactly whether or not the transmission was successful[1]. Packets which were not successfully received are immediately retransmitted.

The performance metric considered in this study is the *lateness probability*, i.e., the probability that a packet is not successfully delivered within D slots of its arrival. In the presence of a smoothing buffer at the receiver, this is the probability that the jitter experienced by a packet is too large to be absorbed by the buffer itself.

3 The channel error process

The packet error process is derived from Clarke's model for the Rayleigh fading channel [19, 20]. It is assumed here for simplicity that during a packet duration the value of the fading envelope does not change significantly[2]. Furthermore, if the value of the envelope is above a certain threshold, the packet is assumed to be correctly decoded with probability 1, and incorrectly otherwise [1]. Long-term propagation effects (path loss and shadowing) are usually compensated by power control and will be ignored here.

The key fading parameters in this context are the *fading margin*, F, and the *normalized Doppler frequency*, $f_D T$, where f_D is the maximum Doppler shift (mobile speed divided by the carrier wavelength) and T is the duration of a packet (slot). The fading margin is the maximum fading attenuation which still allows correct reception, i.e., $F = 10$ dB means that a packet transmission is successful as long as the value of the fading envelope does not drop more than 10 dB below its mean. The normalized Doppler frequency is directly related to the correlation properties of the error process.

The two-state Markov model can be obtained from Clarke's model by matching the average packet error rate and the conditional probability of a success given a failure, which can be either measured from a simulation or computed following Clarke's analytical model [20].

The computation of the lateness probability for the simple system described above and for the two-state error model has been performed analytically in [3]. However, when comparing the analytical results based on the Markov approximation for the Rayleigh fading channel with the performance obtained from its true simulation, significant differences are observed. As shown in Figure 1, the analytical prediction may be accurate for small values of the delay parameter (which is the less interesting case), whereas the match is much worse in the more interesting right-hand side of the graph. This behavior can be explained by looking at the actual statistics of the length of bursts of good and bad slots. For example, Figure 2 shows the complementary probability distribution of the bad burst length for $f_D T = 0.01$ and some typical values of F. On the scale used, a two-state Markov model would result in straight lines, which is obviously not the case here. In fact, the curves appear to exhibit a two-slope behavior, which cannot be correctly captured by a two-state approximation.

---

[1] Extensions to erroneous feedback and finite round-trip are discussed in [3].

[2] This assumption is well verified for all values of the Doppler frequency and data rates considered in this study.

The reason why throughput predictions are accurate is that in evaluating throughput the time span of important events does not extend much beyond the round-trip delay. That is, fitting the channel error process with a Markov process which correctly approximates the bad burst lengths up to just a few slots will provide good results simply because the region where the model is not accurate plays a negligible role in determining the performance. On the other hand, in delay and queueing analysis, in which longer time scales are to be considered, such differences between error model and true behavior may lead to significantly inaccurate performance evaluations.

A look at the statistics of the good burst duration (plotted in Figure 3), on the other hand, reveals that a geometric model for the burst length may be good enough. Evidence in support of this conjecture is given by Figure 4, which reports results of a simulation where the true statistics was used for the bad burst duration, whereas a geometric approximation was used for the good periods. It can be see that the source of error in using a two-state Markov approximation lies in the inaccuracy of the geometric model for the error bad bursts.

4 Higher-order models for the fading channel

The above results demonstrate the need for higher-order error models in applications which require consideration of a larger time scale. This problem can obviously be addressed by increasing the size of the state space of the Markov chain beyond two. However, the beauty of the two-state Markov model does not lie in its accuracy (which is actually not very good in some cases, as just shown), but rather in the ease of handling it analytically when doing performance analysis of protocols running on top of such error processes. The strength of the two-state approach is then the fact that it enables analytical predictions, which make it possible to obtain both quick results and revealing insight.

Based on this point of view, it appears that the straightforward approach of increasing the state space of the chain until some prescribed accuracy of the match is achieved (according to some metric) does not satisfy our main goal, namely, *capturing the essential channel behavior with a minimal number of states.* Therefore, even though accurate channel models can be obtained as in, e.g., [13, 14, 15], we focus in this paper on a much smaller class of models.

Specifically, by again looking at the plots for the burst durations given in Figures 2 and 3, it appears that using just one good channel state may be adequate, whereas a two-slope approximation for the error burst duration would require two bad channel states. In general, we can approximate the complementary distribution of the error burst length, $P[\text{bad burst duration} > x]$, by fitting it with the following function

$$F(x) = \sum_{i=1}^{N} \alpha_i e^{-\beta_i x}, \tag{1}$$

with $\sum_{i=1}^{N} \alpha_i = 1$. Once the parameters of this function have been identified by appropriate fit with the simulation results, on can build a Markov chain with $N+1$ states which has $F(x)$

as the distribution of the bad burst length[3]. A possible choice[4] is a chain where state 0 is the (single) good state, states $1, \ldots, N$ are error states, and the transition matrix is given by

$$P = \begin{pmatrix} p_0 & \alpha_1 q_0 & \alpha_2 q_0 & \cdots & \alpha_N q_0 \\ 1-\beta_1 & \beta_1 & 0 & \cdots & 0 \\ 1-\beta_2 & 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1-\beta_N & 0 & 0 & \cdots & \beta_N \end{pmatrix}, \qquad (2)$$

where $p_0 = 1 - q_0$ can be found by matching the average error rate.

5 Computation of the lateness probability

The delay analysis for the system described in Section 2 has been presented in [3] for the two-state Markov channel model, and is based on the analytical evaluation of the steady-state distribution of the number of packets in the queue and on the recursive evaluation of the number of successful slots in a given time duration. Except for some analytical simplifications which lead to closed-form expressions, the same approach can be applied in this case. In fact, the recursion can be rewritten to account for the increased number of states, whereas the derivation of the steady-state queue length distribution can be obtained in general via matrix-geometric techniques [21], or even in closed form for small N (we will actually focus on $N = 2$ in this study). The interested reader is referred to [3, 21, 22] for more details.

6 Results and discussion

In this section, we present some numerical results obtained based on a three-state Markov approximation for the packet error process induced by the Rayleigh fading. This Markov model has been obtained by fitting the burst length as explained in Section 4, and has been used in the analytical evaluation of the lateness probability. These results are compared to the performance obtained by running the ARQ protocol over the fading channel according to Jakes' simulator [19].

Figure 5 shows the lateness probability for various values of the fading margin and for $f_D T = 0.01$. This corresponds, for example, to the transmission of packets of 1500 bits at a rate of 1 Mbps with carrier frequency 1800 MHz and pedestrian speed. Each plot shows curves for two values of the arrival rate, namely $\lambda = 0.5$ and $\lambda = 0.9$ (other values of λ show similar behavior). For each value of λ, three curves are plotted, i.e., the analytical results for two-state (solid lines) and three-state (dashed lines) Markov models, and the simulated performance (symbols). It is clear that the three-state model provides adequate accuracy, whereas the two-state model may be exceedingly optimistic.

---

[3] We do not address here in detail the issue of the independence of the dwelling times in different states which is inherent in the Markov characterization.

[4] Since a Markov chain with $N+1$ states has $N(N+1)$ independent entries in the transition matrix, whereas the constraints are the parameters of $F(x)$ and the average packet error rate (giving $2N$ equations), it is clear that for $N > 1$ there exists an infinite number of chains which result in the same error statistics.

Other values of the fading rate are considered in Figure 6, where results are plotted for $F = 25$ dB and for $f_D T = 0.002, 0.05$ (to be compared with $f_D T = 0.01$ in Figure 5c). It is seen that the accuracy of the approximation is still satisfactory in most cases, although for some values of the parameters this may not be true, as shown in Figure 7 for $F = 15$ dB and fast fading ($f_D T = 0.05$). In the latter case, the true performance curves change slope twice, and after some value of the delay, the approximate results have the wrong slope and become less and less accurate. Based on what we observed in passing from two to three states, we conjecture that adding another error state may provide the additional degree of freedom necessary to produce that second change of slope. Investigation of this conjecture is left for future study.

We stress the fact that the analytical results presented approximate the simulated performance very closely, which is remarkable given the extreme simplicity of the error model compared to the original Rayleigh fading process. It has been reported in the literature that accurate Markov modeling can be achieved if tens or even hundred of states are used. Here we are getting very good results by using only three!

In order to verify the generality of our findings, we also give another example of application, in which a more complicated protocol is used. Again under the assumption of instantaneous and error-free feedback, we consider an idealized Transport Layer protocol, which can be taken as a representation of TCP. The window adaptation algorithm and the loss recovery mechanism with fast retransmit are as in TCP Tahoe [23]. The protocol is run over a Radio Link Protocol (RLP) which recovers packet errors through retransmission of the erroneous TCP packets. A single connection is assumed, and the TCP packet length is taken to be constant and equal to the radio link packet length.

The radio link protocol is only allowed to transmit a given packet for a finite number of times, after which it will discard the packet and ask the TCP layer to pass down the next one[5]. The discarded packet is therefore considered lost from the TCP viewpoint, and this situation will be handled according to the protocol rules (i.e., if the window size and the channel behavior cause K packets to be successfully received, fast retransmit will occur, otherwise timeout expiration will be waited for). The metric of interest is again the complementary distribution of the packet delay. Bernoulli arrivals are assumed.

Results for $F = 20$ dB and $f_D T = 0.01$ are shown in Figure 8. Again, results obtained by using the two-state and the three-state Markov models are compared with those for the true Rayleigh fading simulation[6]. As before, the two-state model leads to optimistic results, whereas the three-state model gives a better approximation.

From the results of this study, we can draw the following conclusions:

- the assessment of the goodness of an error model must take into account the metric under study, i.e., a model which gives very good approximation for one metric (e.g., throughput) may turn out to be poor for others (e.g., delay)
- even though the error process induced by Rayleigh fading has nontrivial statistical properties and the interaction between this process and the queueing behavior of an ARQ

---

[5] When asking another packet, the RLP does not signal to the TCP whether the previous one was successfully delivered or dropped due to too many retransmissions, thereby complying with layering requirements.

[6] Even for the Markov channel models, the performance shown has been obtained by simulation, since no analytical approach has been developed for the performance analysis of the TCP/RLP stack.

protocol or a TCP/RLP stack is involved, a surprisingly simple channel model has been found to provide very accurate predictions, making analytical computations possible in some cases

- complicated Markov models with many states, which certainly give a better approximation of the error process but may not give as much insight, are not necessarily to be preferred to simple (though a little less accurate) models
- the goodness of the three-state error model depends also on the parameters involved, as we found that for some values of F and $f_D T$ is better than for others; extension of the procedure to more states seems straightforward, and such a study is left as future work.

References

[1] M. Zorzi, R.R. Rao, L.B. Milstein, "On the accuracy of a first-order Markov model for data block transmission on fading channels," in *Proc. IEEE ICUPC'95*, pp. 211-215, Nov. 1995.

[2] M. Zorzi, R.R. Rao, L.B. Milstein, "ARQ error control on fading mobile radio channels," *IEEE Trans. Veh. Tech.*, vol. 46, pp. 445–455, May 1997.

[3] M. Zorzi, R.R. Rao, "ARQ Error Control for Delay-Constrained Communications on Short-Range Burst-Error Channels," in *Proc. VTC'97*, pp. 1528–32, May 1997. Also, under review for *IEEE Trans. Comm.*

[4] M. Zorzi, R.R. Rao, "On the Statistics of Block Errors in Bursty Channels," *IEEE Trans. Comm.*, vol. 45, pp. 660–667, Jun. 1997.

[5] M. Zorzi, R.R. Rao, "The Role of Error Correlations in the Design of Protocols for Packet Switched Services," in *Proc. 35th Annual Allerton Conference*, pp. 749–758, Sep. 1997.

[6] M. Zorzi, R.R. Rao, "Energy constrained error control for wireless channels," *IEEE Personal Communications Magazine*, vol. 4, pp. 27–33, Dec. 1997.

[7] W. Turin, *Performance analysis of digital transmission systems*, Computer Science Press, NY, 1990.

[8] E.N. Gilbert, "Capacity of a burst-noise channel", *Bell System Tech. Journal*, vol. 39, pp. 1253–1266, Sept. 1960.

[9] F. Swarts, H.C. Ferreira, "Markov characterization of channels with soft decision outputs," *IEEE Trans. Commun.*, vol. COM-41, pp. 678-682, May 1993.

[10] H.S. Wang, "On verifying the first-order Markovian assumption for a Rayleigh fading channel model," in *Proc. ICUPC'94*, pp. 160-164, San Diego, CA, Sep. 1994. Also in *IEEE Trans. Veh. Tech.*, vol. VT-45, pp. 353–357, May. 1996.

[11] H.S. Wang, N. Moayeri, "Finite-state Markov channel — a useful model for radio communication channels", *IEEE Trans. Veh. Technol.*, vol. VT-44, pp. 163-171, Feb. 1995.

[12] M. Zorzi, "Outage and Error Events in Bursty Channels," *IEEE Trans. Comm.*, vol. 46, pp. 349–356, Mar. 1998.

[13] Q. Zhang, S.A. Kassam, "Finite-state Markov model for Rayleigh fading channels," submitted for publication.

[14] W. Turin, R. Van Nobelen, "Hidden Markov modeling of fading channels," in *Proc. VTC98*, pp. 1234–38, May 1998.

[15] Y. Y. Kim, S.Q. Li, "Modeling fast fading channel dynamics for packet data performance analysis," in *Proc. INFOCOM98*, pp. 1292–1300, Apr. 1998.

[16] B. D. Fritchman, "A binary channel characterization using partitioned Markov chains," *IEEE Trans. on Info. Theory*, vol. IT-13, Apr. 1967.

[17] A. M. Chen, Ramesh R. Rao, "On Tractable Wireless Channel Models," in *Proc. PIMRC98*, Paper ID inv017, Boston, MA, September. 1998.

[18] M. Zorzi, A. Chockalingam, R.R. Rao, "Performance Analysis of TCP on Channels with Memory," submitted for publication.

[19] W.C. Jakes, Jr., *Microwave mobile communications*, New York: John Wiley & Sons, 1974.

[20] R.H. Clarke, "A statistical theory of mobile radio reception", *Bell System Tech. Journal*, vol. 47, pp. 957-1000, Jul. 1968.

[21] M.F. Neuts, *Matrix-geometric solutions in stochastic models: an algorithmic approach*, Baltimore: Johns Hopkins University Press, 1981.

[22] R.A. Howard, *Dynamic probabilistic systems*, John Wiley & Sons, 1971.

[23] W. R. Stevens, *TCP/IP Illustrated, Volume 1*, Addison Wesley, 1994.

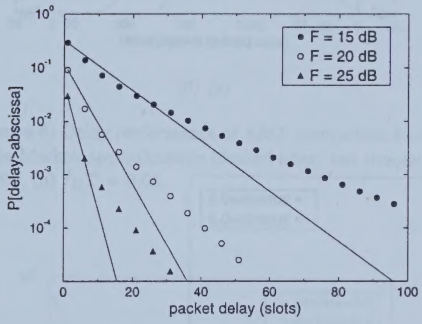


Figure 1: Delay performance of ARQ: comparison between two-state Markov approximation and simulated fading process, $F = 15, 20, 25$ dB, $f_D T = 0.01, \lambda = 0.9$.

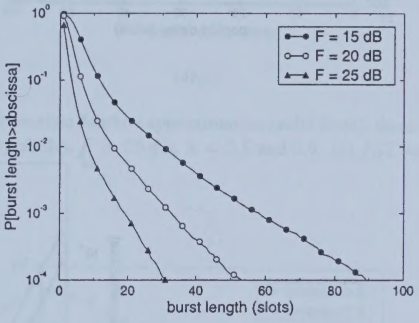


Figure 2: Bad burst length distribution (simulated) for $F = 15, 20, 25$ dB and $f_D T = 0.01$.

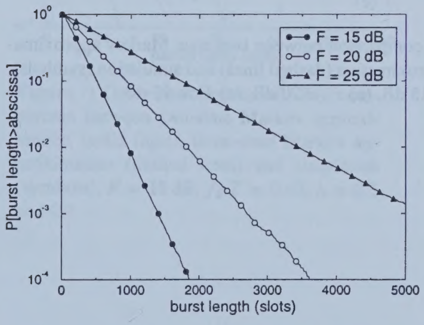


Figure 3: Good burst length distribution (simulated) for $F = 15, 20, 25$ dB and $f_D T = 0.01$.

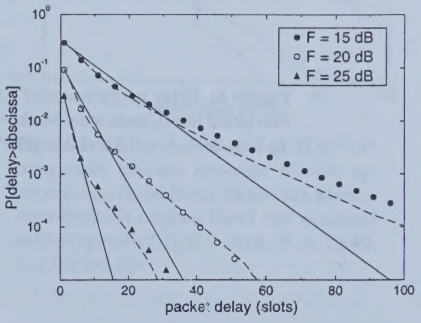
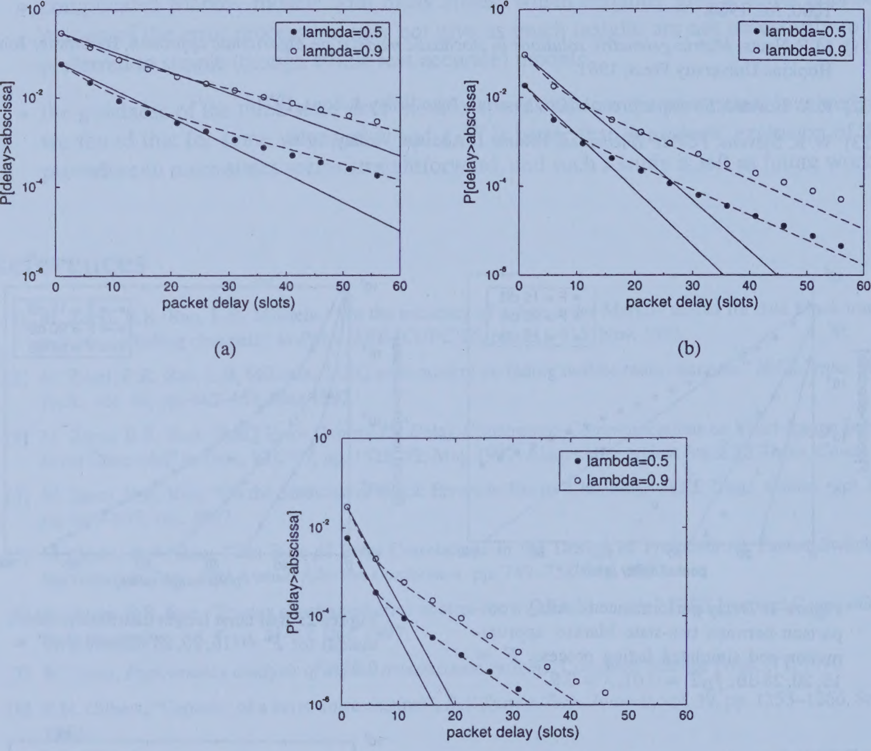


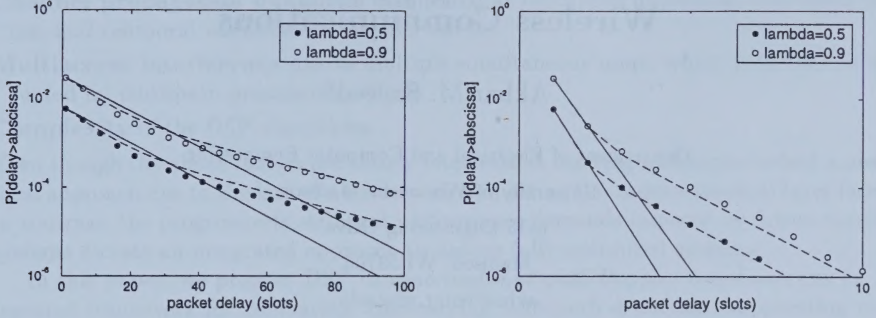
Figure 4: Delay performance of ARQ: comparison between two-state Markov approximation (solid lines), combination of exponential approximation of good bursts and true distribution of bad bursts (dashed lines), and simulated fading process (symbols), $F = 15, 20, 25$ dB, $f_D T = 0.01, \lambda = 0.9$.



(a) (b)

(c)

Figure 5: Delay performance of ARQ: comparison between two-state Markov approximation (solid lines), three-state Markov approximation (dashed lines) and simulation (symbols), $f_D T = 0.01$, $\lambda = 0.5$ and 0.9. (a) $F = 15$ dB, (b) $F = 20$ dB, (c) $F = 25$ dB.



(a) (b)

Figure 6: Delay performance of ARQ: comparison between two-state Markov approximation (solid lines), three-state Markov approximation (dashed lines) and simulation (symbols), $F = 25$ dB, $\lambda = 0.5$ and 0.9. (a) $f_D T = 0.002$, (b) $f_D T = 0.05$.

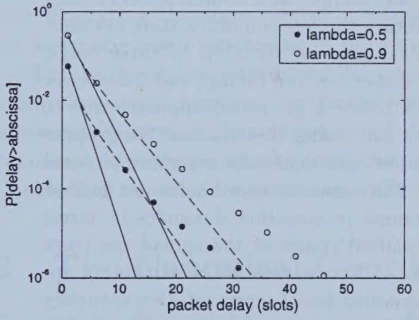


Figure 7: Delay performance of ARQ: comparison between two-state Markov approximation (solid lines), three-state Markov approximation (dashed lines) and simulation (symbols), $F = 15$ dB, $f_D T = 0.05$, $\lambda = 0.5$ and 0.9.

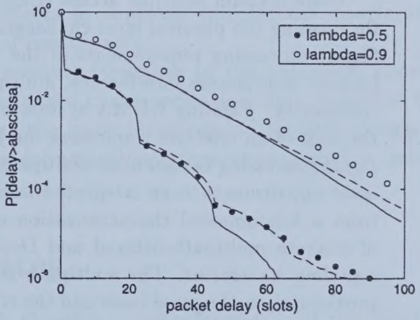


Figure 8: Delay performance of TCP/RLP: comparison between two-state Markov approximation (solid lines), three-state Markov approximation (dashed lines) and simulated fading (symbols), $f_D T = 0.01$, $F = 20$ dB, $\lambda = 0.5$ and 0.9.

Canonical Multipath-Doppler Coordinates in Wireless Communications

Akbar M. Sayeed*

Department of Electrical and Computer Engineering
University of Wisconsin–Madison
1415 Engineering Drive
Madison, WI 53706
akbar@engr.wisc.edu
Tel: (608) 265-4731, Fax: (608) 262-1267
http://www.engr.wisc.edu/ece/

Abstract

Code division multiple access (CDMA) has emerged as a dominant technology for meeting the physical layer challenges of future wireless communication systems. Signal processing requirements in the physical layer are dictated by three major factors: multiaccess interference, multipath dispersion and fading, and transceiver complexity. Existing CDMA system designs reflect a piecemeal approach due to the lack of an effective framework for jointly addressing these issues. We propose signal processing in canonical multipath-Doppler coordinates for attacking physical layer impairments in an integrated fashion. The canonical coordinates are derived from a fundamental characterization of channel propagation dynamics in terms of discrete multipath-delayed and Doppler-shifted copies of the spread-spectrum signaling waveforms. The multipath-Doppler shifted waveforms constitute an approximately orthogonal basis and the corresponding signal representation naturally connects the various channel effects. First, all processing relating to multipath propagation can be directly performed in the canonical coordinates. Second, the same coordinates provide a canonical subspace-based representation of the desired signal and interference which fully incorporates channel dispersion effects. Finally, the maximally parsimonious nature of the coordinates and their simple computation afford a direct handle on transceiver complexity. Various facets of the integrated framework are illustrated in the context of interference suppression, channel estimation, and diversity processing.

1 Introduction

Code division multiple access (CDMA) has emerged as one of the most promising technologies for meeting the challenges of the physical network layer, as evident from its prominence in the wireless infrastructure envisioned for the third and future generation

---

*This research was supported in part by the Wisconsin Alumni Research Foundation.

systems. Innovative DSP techniques are playing a progressively important role in the design of CDMA systems due to the high complexity of processing required in the physical layer. The major DSP challenges in CDMA system design stem from three key factors that have a significant impact on performance:

Channel propagation dynamics manifested as multipath dispersion, multipath fading, and temporal variations or Doppler effects.

Multiaccess interference due to multiple simultaneous users, which is further accentuated by multipath propagation effects.

Complexity of the DSP algorithms.

Even though the above factors are clearly interrelated, existing techniques reflect a piecemeal approach due to the lack of a framework connecting the various physical layer facets. In contrast, the progressively stringent performance demands imposed by future wireless systems dictate an integrated approach to deliver fully optimized systems.

In this paper, we propose DSP in *canonical multipath-Doppler coordinates* as an integrated framework for combating time-varying multipath distortion, suppressing multiaccess interference, and managing the complexity of the resulting algorithms. The canonical coordinates are derived from a fundamental characterization of the channel propagation dynamics in terms of discrete multipath-delayed and Doppler-shifted copies of the *transmitted* spread-spectrum waveform which constitute a *fixed* canonical basis for representing the received signal [1]. A well-known advantage of CDMA, by virtue of spread-spectrum signaling, is its remarkable ability to combat fading by *exploiting* multipath propagation effects [2]. Our framework promises a new DSP innovation: further exploiting the time-varying multipath effects to suppress multiaccess interference via subspace-based processing in the canonical coordinates.

The next section introduces the generic receiver structure in the canonical coordinates. Section 2 describes the fundamental channel characterization that underlies the proposed framework. Sections 4 and 5 discuss the basic ideas behind receiver design in canonical coordinates, emphasizing channel modeling and interference suppression issues. Concluding remarks are presented in Section 6.

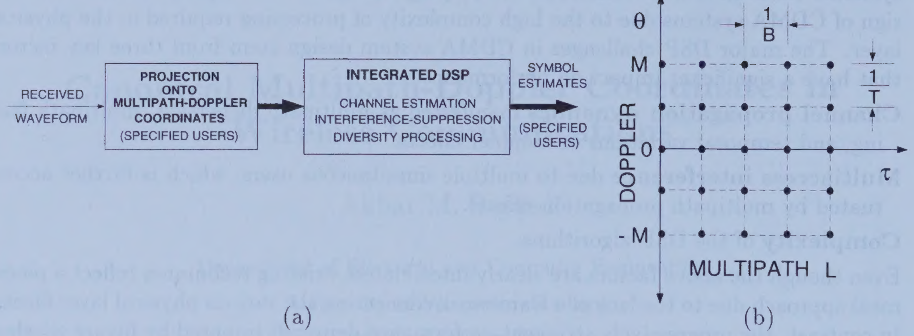
2 The Canonical Coordinates

Figure 1 describes *DSP in canonical multipath-Doppler coordinates* that is at the heart of our framework. The front-end processing corresponds to projecting the received waveform onto the canonical coordinates. Each user corresponds to unique coordinates defined by its spreading code, which are computed for each received symbol. Canonical coordinates, taken together for all users and symbols of interest, constitute sufficient statistics for demodulation—*all DSP can be performed in the canonical coordinates.*

The coordinates for each symbol of a particular user are computed by correlating the received signal with time- and frequency-shifted copies of the user spreading waveform

$$\begin{aligned} z_{ml} &= \int r(t) q^*\left(t - \tau_o - \frac{l}{B}\right) e^{-j\frac{2\pi m t}{T}}\, dt, \qquad (1) \\ l = 0, \cdots, L \quad &; \quad m = -M, \cdots, 0, \cdots, M, \end{aligned}$$

where $r(t)$ denotes the received waveform, τ_o denotes the user delay, and $q(t)$ is the user spreading waveform of duration T and bandwidth B. In effect, the canonical multipath-Doppler coordinates define an $(L+1)(2M+1)$-dimensional subspace for representing the


RECEIVED
WAVEFORM
PROJECTION
ONTO
MULTIPATH-DOPPLER
COORDINATES
(SPECIFIED USERS)
INTEGRATED DSP
CHANNEL ESTIMATION
INTERFERENCE-SUPPRESSION
DIVERSITY PROCESSING
SYMBOL
ESTIMATES
(SPECIFIED
USERS)
θ
1/B
1/T
M
0
- M
DOPPLER
τ
MULTIPATH L

(a) (b)

Figure 1: DSP at the receiver in canonical multipath-Doppler coordinates. (a) Overall receiver schematic. (b) The uniformly spaced canonical coordinate grid.

channel distorted signal of each user. The subspace is spanned by the *fixed* basis

$$q_{ml}(t) = q\left(t - \frac{l}{B}\right) e^{j\frac{2\pi m t}{T}}, \tag{2}$$

determined by the user's spreading waveform $q(t)$ [1]. We note that under certain conditions $\{q_{ml}(t)\}$ constitute an approximately orthogonal basis [1].

As illustrated in Figure 1(b), the indices l and m correspond to a uniform sampling of multipath delays τ ($\Delta\tau = 1/B$) and Doppler shifts θ ($\Delta\theta = 1/T$), respectively, that are encountered during propagation. The front-end projection onto the canonical coordinates can be performed efficiently via a time-frequency generalization of the RAKE receiver — a bank of RAKE receivers matched to different canonical Doppler frequencies [1].

As we will see in subsequent sections, in addition to providing an equivalent, robust characterization of virtually all existing channel modeling approaches, canonical multipath-Doppler coordinates connect the channel propagation and multiaccess interference effects in a very natural and intuitive fashion which leads to new insights and methodologies. First of all, by their very nature, processing relating to channel propagation can be directly performed in the canonical coordinates [1]. Moreover, the same coordinates provide a natural subspace-based representation of the desired signal and multiaccess interference which fully incorporates channel distortion effects and leads to new techniques for interference suppression. Finally, the parsimonious nature of the canonical coordinates affords a direct handle on DSP complexity.

3 The Canonical Signal Representation

In this section, we describe a fundamental characterization of channel propagation effects that underlies canonical multipath-Doppler coordinates [1]. Propagation effects can be generally modeled by a linear time-varying system [3, 2] and are best illustrated by focusing on transmission of a single symbol waveform $q(t)$ in a single-user system:

$$r(t) \;=\; s(t) + n(t), \tag{3}$$

$$s(t) \;=\; \int_0^{T_m}\int_{-B_d}^{B_d} H(\theta,\tau)q(t-\tau)e^{j2\pi\theta t}d\theta d\tau, \quad 0 \le t \le T, \tag{4}$$

where $r(t)$ is the received noisy waveform consisting of the information bearing signal $s(t)$ and additive white Gaussian noise (AWGN) $n(t)$. Channel propagation is characterized by the multipath-Doppler *spreading function* $H(\theta,\tau)$ which depicts the temporal and spectral dispersion produced by the channel.[1] The variable τ corresponds to multipath propagation delays with T_m denoting the *multipath spread* of the channel. Doppler shifts are denoted by θ, with B_d denoting the *Doppler spread* of the channel.

The notion of canonical coordinates is based on the following basic idea: due to the inherently finite duration T and essentially finite bandwidth B of the transmitted waveform, the receiver "sees" only *finitely many* degrees of freedom in the signal [3, 4, 1]. These essential degrees of freedom are captured by the canonical multipath-Doppler coordinates as dictated by the following fundamental characterization [1]

$$s(t) \approx \frac{1}{TB}\sum_{l=0}^{L}\sum_{m=-M}^{M} \widetilde{H}\left(\frac{m}{T},\frac{l}{B}\right) q\left(t-\frac{l}{B}\right) e^{j\frac{2\pi m t}{T}}, \quad 0 \le t \le T, \tag{5}$$

which leads to the canonical basis (2) corresponding to the uniform grid in Figure 1(b). The coefficients in the basis expansion are samples of the smoothed spreading function

$$\begin{aligned}\widetilde{H}(\theta,\tau) \;&=\; TB\int_0^{T_m}\int_{-B_d}^{B_d} H(\theta',\tau') \\ &\qquad \mathrm{sinc}((\theta-\theta')T)\mathrm{sinc}((\tau-\tau')B)d\theta' d\tau' \end{aligned} \tag{6}$$

that arises due to the time limited and essentially band limited nature of $q(t)$ [1]. The approximation in (5) can be made arbitrarily precise by increasing the number of summation terms — in practice, only a few dominant terms capture virtually all the signal energy. The number L of dominant multipath coordinates is equal to the multipath spread normalized by the temporal signaling resolution $(1/B)$: $L = \lceil T_m B \rceil$. Similarly, the number of dominant Doppler coordinates is equal to the Doppler spread normalized by the frequency resolution $(1/T)$: $M = \lceil T B_d \rceil$.

We note that the definition of the effective bandwidth B affects the structure of the basis (2) and the accuracy of (5). In particular, for direct sequence spread-spectrum waveforms, B is inversely related to the chip duration $T_c = T/N$, where N is the spreading gain. The choice $B \approx 1/T_c$ yields an approximately orthogonal basis (2) at the expense of some loss of accuracy in (5). The accuracy of the representation (2) can be improved via the choice $B \approx \mathcal{O}/T_c$, $\mathcal{O} \ge 2$ (oversampling by a factor of $\mathcal{O}$, typically 2 or 4), albeit at the cost of losing orthogonality of the basis (2) (more discussion in Section 4.1).

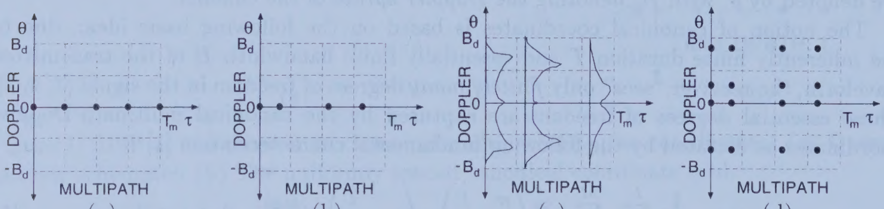
4 Channel Modeling in Canonical Coordinates

The characterization (5) asserts that all information about the channel propagation dynamics that is observable at the receiver is captured by the canonical coefficients $\widetilde{H}_{ml} = \widetilde{H}(m/T, l/B)$ on the uniform multipath-Doppler grid in Figure 1(b). The reason is that due to the essentially finite time-bandwidth product of the transmitted waveform

---

[1] We note that T may be significantly longer than the *intersymbol* duration in order to exploit channel diversity, thereby making Doppler effects significant over the symbol duration [1].

$q(t)$, the received signal $s(t) = \alpha q(t-\tau)e^{j2\pi\theta t}$ corresponding to arbitrary (θ, τ) coordinates admits an effective representation with respect to the canonical basis (2). Figure 2 illustrates several important examples of existing channel models that are captured by modeling in canonical coordinates. In contrast to modeling via a *fixed* basis in the canonical coordinates, most existing models require knowledge of the true multipath delays and Doppler shifts which leads to *data-dependent* basis waveforms and hence increased complexity.



(a) (b) (c) (d)

Figure 2: Illustration of various channel models captured by the canonical coordinates via (5). (a) The tapped delay-line multipath model used by the RAKE receiver [2]. (b) Modeling of arbitrary multipath delays [5]. (c) Time-varying extension of (b) via autoregressive modeling of the Doppler spectrum [6, 7]. (d) Time-varying extension of (b) via arbitrary discrete Doppler shifts [8].

We note that stochastic modeling of $H(\theta, \tau)$ can also be efficiently accomplished by imposing an appropriate statistical model on the (few) coefficients $\widetilde{H}_{ml}$. Statistical modeling is appropriate for capturing the phenomenon of fading. For example, in the wide-sense stationary uncorrelated scattering (WSSUS) model, $H(\theta, \tau)$ is modeled as a family of uncorrelated Gaussian random variables [3, 2]. WSSUS modeling imposes a corresponding Gaussian structure on the coefficients $\widetilde{H}_{ml}$ [1].

4.1 Basic Receiver Structure and Channel Estimation

We illustrate the basic receiver structure for binary antipodal signaling. Generalizations to other forms of signaling, such as M-ary or orthogonal signaling, are also possible [1]. If reliable estimates of the channel coefficients $\widehat{H}_{ml}$ are available, such as through a high-SNR pilot transmission, the coherent receiver in the canonical coordinates takes the form

$$\hat{b} = \text{sign}\left\{\text{real}\left(\mathbf{h}^H\mathbf{z}\right)\right\} = \text{sign}\left\{\text{real}\left(\sum_{l=0}^{L}\sum_{m=-M}^{M}\widetilde{H}^*_{ml}z_{ml}\right)\right\} \qquad (7)$$

where $\mathbf{h} = [\widehat{H}_{ml}]$ denotes an $(L+1)(2M+1)$-dimensional vector of channel coefficients, and $\mathbf{z} = [z_{ml}]$ is a vector of projection onto the canonical coordinates in (1).[2] Note that the above canonical receiver does not require estimation of actual multipath delays or Doppler frequencies of dominant scatterers — only synchronization to the "global" delay and Doppler shift, corresponding to the $(0,0)$-th coordinate in the uniform grid in Figure 1(b), is needed, which is required by all receivers. Furthermore, this robustness in channel modeling does not entail a loss in performance, as illustrated next.

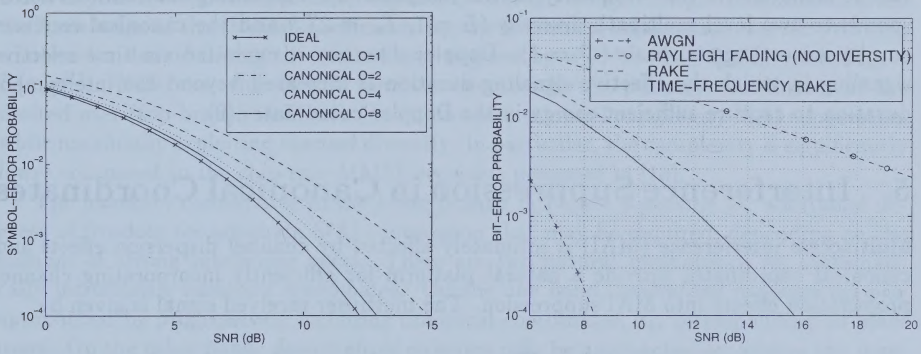
---

[2] $\mathbf{z}^H$ denotes Hermitian transpose of $\mathbf{z}$.

Figure 3(a) demonstrates that the coherent matched filter (MF) receiver (7) in the canonical coordinates delivers near-optimal performance at significantly lower complexity compared to an ideal receiver that requires knowledge of all the dominant multipath delays and Doppler shifts. The comparison is made for a system with $N = 31$ spreading gain and employing a noise-free pilot transmission for channel estimation. A slow fading environment is modeled via $L_T = 64$ uniformly spaced independent multipath components, spread over $T_m = 2T_c$ and all of equal power. The ideal receiver is matched to all the L_T multipaths and takes the form

$$\hat{b} = \text{sign}\left\{\text{real}\left(\sum_{l=0}^{L_T} H^*(\tau_l)\int r(t)q^*(t-\tau_l)dt\right)\right\}. \qquad (8)$$

On the other hand, the canonical receiver is given by (7) corresponding to $M = 0$ (slow fading) and matched to the $L = \lceil T_m/B \rceil = T_m\mathcal{O}/T_c = 2\mathcal{O}$ canonical basis waveforms $\{q(t - lT_c/\mathcal{O})\}$, where $\mathcal{O}$ denotes the oversampling factor.



(a) (b)

Figure 3: (a) Performance of matched filtering in the ideal versus canonical coordinates under slow fading. (b) Joint multipath-Doppler diversity gains achieved by the coherent time-frequency RAKE receiver ($\mathcal{O} = 1$) under typical fast fading scenarios. Theoretical comparison with a 2-path (L=1) coherent RAKE receiver operating under ideal slow fading conditions. The time-frequency RAKE receiver corresponds to only 10% of the total signal energy in $m = \pm 1$ Doppler coordinates that may be attained via time-selective signaling [9].

It is evident from Figure 3(a) that the performance of the canonical receiver for $\mathcal{O} = 4$ or 8 is within 1dB of ideal performance. In addition, the complexity of the canonical receiver is substantially lower than that of the ideal receiver. As opposed to $L+1$ (at most 17 in Figure 3(a)) correlators in the canonical receiver,[3] the ideal receiver requires L_T (64 in this case) correlators. Furthermore, after synchronization, the canonical receiver only requires estimates of $L + 1$ coefficients $\widetilde{H}_l$, whereas the ideal receiver requires estimation of L_T delays τ_l and $H(\tau_l)$. If a pilot channel is available, the canonical channel coefficients

---

[3] Which can be efficiently computed via a tapped delay-line RAKE structure [1].

$\widetilde{H}_l$ can be efficiently computed by projecting the pilot signal onto the canonical subspace spanned by the basis waveforms (2). As noted earlier, most existing receivers espouse the "ideal" structure (8) for dominant scatterers, as depicted in Figure 2(b).

4.2 Diversity Processing

The WSSUS channel affords intrinsic diversity to combat fading that can be exploited via spread-spectrum signaling. Canonical coordinates fully capture the inherent channel diversity that is achieved via the canonical time-frequency RAKE receiver — each coordinate corresponds to a diversity channel [1]. The effective level of diversity is governed by the number of nonvanishing eigenvalues of the signal correlation matrix and corresponds to the number of canonical coordinates. For example, the underlying channel in Figure 3(a) affords roughly 4 level multipath diversity that is efficiently captured by the 5-9 multipath coordinates ($\mathcal{O}$= 2 or 4) in the canonical receiver. Furthermore, the canonical coordinate framework also facilitates exploitation of typical Doppler spreads encountered in fast fading scenarios for enhanced diversity gains. Figure 3(b) depicts realizable joint multipath-Doppler diversity gains under practical fast fading scenarios that can be achieved via time-frequency RAKE reception [1]. The fading environment corresponds to two level multipath diversity ($L = 1$, $T_m = 2T_c$) and the canonical receivers employ an orthogonal basis ($\mathcal{O} = 1$). Doppler diversity is exploited via time-selective signaling in which the effective signaling duration is increased beyond the intersymbol duration to capture sufficient energy in the Doppler coordinates [9].

5 Interference Suppression in Canonical Coordinates

Multiaccess interference (MAI) is intimately affected by channel dispersion effects and canonical coordinates provide a natural platform for efficiently incorporating channel propagation effects into MAI suppression. The multiuser received signal is given by

$$r(t) \;=\; \sum_{k=1}^{K} b_k s_k(t-\tau_k) + n(t), \tag{9}$$

$$s_k(t) \;\approx\; \frac{1}{TB}\sum_{l=0}^{L_k}\sum_{m=-M_k}^{M_k} \widetilde{H}_{k,ml}\, q_{k,ml}(t). \tag{10}$$

and the k-th user is associated with a $(L_k+1)(2M_k+1)$-dimensional canonical subspace defined by the waveforms $q_{k,ml}(t)$ in terms of its spreading waveform $q_k(t)$ via (2). This formulation provides a new subspace-based approach for suppressing MAI by exploiting the additional degrees of freedom afforded by channel dispersion effects.

The basic idea is illustrated in Figure 4 depicting the special case of coherent linear processing for demodulating the desired user, say user 1. The top branch in Figure 4 corresponds to the single-user MF receiver, denoted by $\mathbf{w}_p$, which combines the *primary* coordinates $\mathbf{z}_1$ — the $D_p = (L_1+1)(2M_1+1)$ *active* multipath-Doppler coordinates — carrying signal energy of the desired user (user 1)

$$\mathbf{z}_1 = b_1\mathbf{Q}_{11}\mathbf{h}_1 + \sum_{k=2}^{K} b_k\mathbf{Q}_{1k}(\tau_1,\tau_k)\mathbf{h}_k + \mathbf{n}_1, \tag{11}$$

where $\mathbf{Q}_{jk}(\tau_j,\tau_k)$ denotes cross correlation between the basis waveforms $\{q_{j,ml}(t-\tau_j)\}$ and $\{q_{k,ml}(t-\tau_k)\}$ of users j and k, respectively. The maximal-ratio-combiner (MRC) which

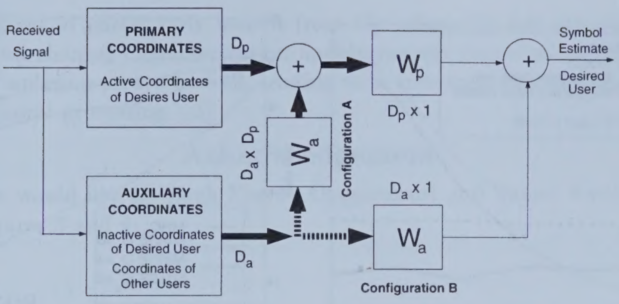


Figure 4: Generic linear receiver structure.

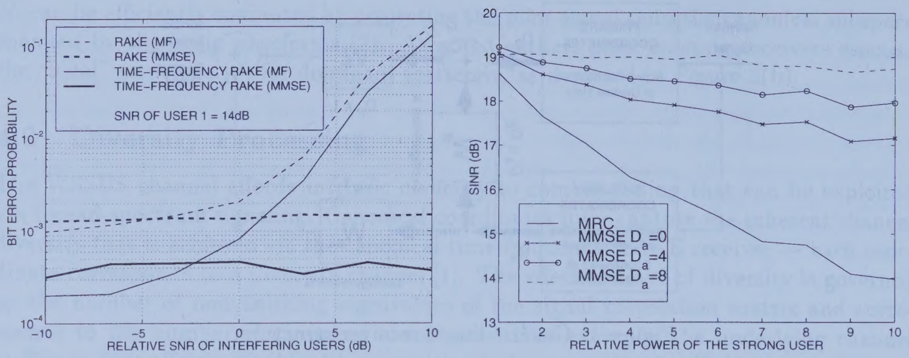
exploits multipath-Doppler diversity but ignores MAI is given by $\mathbf{w}_p = \mathbf{h}_1 = [\widetilde{H}_{1,ml}]$ (same as (7)). However, by virtue of the subspace structure, even the single-user MF receiver possesses MAI suppression capability through design of $\mathbf{w}_p$. For example, given the knowledge of the channel coefficients $\mathbf{h}_1$, the minimum-mean-squared-error (MMSE) solution for demodulating the desired user is given by $\mathbf{w}_p^{mmse} = \mathbf{R}_{\mathbf{z}_1}^{-1}\mathbf{Q}_{11}\mathbf{h}_1$ where $\mathbf{R}_{\mathbf{z}_1} = \mathrm{E}\{\mathbf{z}_1\mathbf{z}_1^H\}$. Note that this MMSE receiver operates in the primary coordinates of the desired user (top branch in Figure 4) and provides minimal complexity MAI suppression while maximally exploiting channel diversity. In particular, the complexity is significantly lower compared to the chip-rate MMSE receivers proposed in [10].

The *auxiliary* coordinates in the lower branch in Figure 4 provide D_a additional degrees of freedom for adequate MAI suppression that may be required, depending on the number of strong interfering users. There are two natural mechanisms for systematically introducing auxiliary coordinates. On the one hand, *centralized* receivers may be approached by progressively including canonical coordinates, $\mathbf{z}_k$, corresponding to *other* users. On the other hand, *decentralized* receivers may be approached by adding the *inactive* canonical coordinates of the desired user (outside the $(L_1+1)(2M_1+1)$-dimensional *active* square in Figure 1(b)). In between the two extremes is a whole range of *hybrid* receivers that attain desired auxiliary degrees of freedom by combining decentralized and centralized coordinates. Another dimension for flexibility comes from the various ways in which the primary and auxiliary coordinates are *combined*, as depicted by Configurations A and B in Figure 4. Configuration A corresponds to MAI suppression, via $\mathbf{w}_a$, followed by diversity combining/MAI suppression through $\mathbf{w}_p$. Configuration B corresponds to *joint* diversity combining and MAI suppression through choice of $\mathbf{w}_a$ and $\mathbf{w}_p$.

Figure 5 depicts the performance of MAI-resistant linear coherent receivers in the canonical coordinates (oversampling factor $\mathcal{O} = 1$; $K = 4$ users). A centralized MMSE multiuser detector which uses coordinates of all 4 interfering users as auxiliary coordinates is depicted in Figure 5(a) [11]. The fading conditions are the same as in Figure 3(b). It generalizes the conventional MMSE multiuser detector to fully incorporate dispersion effects and to maximally exploit channel diversity [1]. Figure 5(b) depicts the signal-to-interference-and-noise ratio (SINR) of a decentralized MMSE receiver as a function of the relative interference power for different numbers D_a of auxiliary multipath coordinates.[4]

---

[4] SNR of desired user = 20dB.



(a) (b)

Figure 5: MAI suppression ($K = 4$ users) via linear coherent processing in canonical coordinates ($\mathcal{O} = 1$). (a) Near-far resistance of a (maximal complexity) MMSE receiver with *centralized* auxiliary coordinates corresponding to *all* the interfering users. Multipath fading conditions same as in Figure 3(b). The joint multipath-Doppler diversity gains of the time-frequency RAKE receiver are directly inherited by the MMSE receiver (see also Figure 3(b)). The MF receivers correspond to single-user MRCs. (b) SINR of a (minimal complexity) MMSE receiver with increasing number D_a of *decentralized* auxiliary coordinates. Slow fading conditions with 4 multipaths ($L = 3$).

It is based on a slow fading environment ($T_m = 4T_c$; $N = 64$) with 4 users, one of which is a strong interferer. Figure 5(b) illustrates that even minimal complexity MAI suppression in the primary coordinates can yield significant gains when there are a few strong interferers.

We note in passing that MAI affects channel estimation much the same way as it affects data demodulation. Thus, the above general ideas apply to MAI-resistant channel estimation as well. For example, in centralized receivers corresponding to Configuration A in Figure 4, the inputs to the primary filter $\mathbf{w}_p$ yield an MAI-resistant estimate of the channel coefficients $\mathbf{h}_1$ of the desired user [11]. Furthermore, the subspace structure afforded by the canonical coordinates can also be leveraged into noncoherent techniques for MAI-resistant timing acquisition and demodulation [12].

6 Conclusion

We have presented the basis ideas behind a new framework for receiver design in canonical multipath-Doppler coordinates that provides a natural platform for dealing with channel propagation effects, multiaccess interference, and the complexity of the resulting algorithms in an integrated fashion. From a channel modeling and estimation perspective, the canonical coordinates obviate the need for estimating arbitrary delays and Doppler shifts of dominant scatterers, thereby substantially reducing receiver complexity. In the context of MAI suppression, the canonical subspace structure provides a systematic approach for designing minimal complexity receivers that in fact exploit propagation effects for MAI mitigation. Adaptive or blind formulations of the hybrid receivers suggested by the framework would particularly benefit from the inherently low complexity. Finally, the fundamental channel characterization underlying the canonical coordinates can be generalized to antenna arrays as well, leading to a corresponding canonical framework for spatio-temporal processing [13].

Acknowledgement

The author would like to thank Eko N. Onggosanusi and Tamer Kadous for help in generating Figures 3 and 4.

References

[1] A. M. Sayeed and B. Aazhang, "Joint multipath-Doppler diversity in mobile wireless communications," *to appear in the IEEE Trans. Commun.*, 1998.

[2] J. G. Proakis, *Digitial Communications.* New York: McGraw Hill, 3rd ed., 1995.

[3] P. A. Bello, "Characterization of randomly time-variant linear channels," *IEEE Trans. Commun. Syst.*, vol. CS-11, pp. 360–393, 1963.

[4] R. Kennedy, *Fading Dispersive Communication Channels.* Wiley, 1969.

[5] A. J. Paulraj and C. B. Papadias, "Space-time processing for wireless communications," *IEEE Signal Processing Magazine*, pp. 49i–83, Nov. 1997.

[6] A. W. Fuxjaeger and R. A. Iltis, "Adaptive parameter esimation using parallel Kalman filtering for spread spectrum code and Doppler tracking," *IEEE Trans. Commun.*, vol. 42, pp. 2227–2230, June 1994.

[7] M. K. Tsatsanis, G. B. Giannakis, and G. Zhou, "Estimation and equalization of fading channels with random coefficients," *Signal Processing*, vol. 53, pp. 211–229, 1996.

[8] G. B. Giannakis and C. Tepedelenlioglu, "Basis expansion models and diversity techniques for blind identification and equalization of time-varying channels," *submitted to Proc. IEEE (preprint)*, Aug. 1997.

[9] S. Bhasyam, A. M. Sayeed, and B. Aazhang, "Time-selective signaling and reception for multipath fading channels," *presented at the 1998 IEEE Intl. Symp. Inform. Th.*, 1998.

[10] U. Madhow and M. L. Honig, "MMSE interference suppression for direct-sequence spread-spectrum CDMA," *IEEE Trans. Commun.*, vol. 42, pp. 3178–3188, Dec. 1994.

[11] A. M. Sayeed, A. Sendonaris, and B. Aazhang, "Multiuser detection in fast fading multipath environments," *to appear in the IEEE J Select. Areas Commun.*, 1998.

[12] A. M. Sayeed and B. Aazhang, "Multiuser timing acquisition over multipath fading channels," in *Proc. 1998 Conf. Inf. Sci. Syst. (CISS'98)*, 1998.

[13] E. N. Onggosanusi, A. M. Sayeed, and B. D. Van Veen, "Canonical space-time processing in CDMA systems," in *Proc. 1999 IEEE Int. Conf. on Acoust., Speech and Signal Proc. — ICASSP '99.*

A New Framework for Power Control in Wireless Data Networks: Games, Utility and Pricing*

David Famolari and Narayan Mandayam and David Goodman
Wireless Information Laboratory (WINLAB)
Dept. of Electrical Engineering and Computer Engineering
Rutgers University
Piscataway, NJ 08854
fam,narayan,dgoodman@winlab.rutgers.edu

Abstract

This work investigates a distributed power control scheme for wireless data based on the economic principles of utility and pricing. Utility is defined as the measure of satisfaction that a user derives from accessing the wireless data network. Properties of utility functions are introduced and a specific function, based on throughput per terminal battery lifetime including forward error control, is presented and shown to conform to those properties. Users enter into a non-cooperative game to maximize their individual utilities by adjusting their transmitter powers. A unique solution to the above game is shown to exist but is not Pareto efficient. A pricing function is then introduced which leads to Pareto improvements for the non-cooperative game. In both cases forward error control can be shown to increase the utility of every user at a lower transmitter power and lead to larger feasible user populations.

1 Introduction

Power control for wireless data based on utility and pricing has been considered in [1]. In that work a utility function was developed and distributed power control was formulated as a non-cooperative game. It was shown that an equilibrium solution to that game existed and was unique, however it failed to be Pareto efficient. Pricing was then introduced and shown to yield Pareto improvements with respect to the utility only game.

In this work we extend the previous results to include the effects of forward error control in the form of error detection and error correction. We show that a utility function based on forward error control can yield a unique solution to the non-cooperative power control game and provide a dramatic increase in user utility and a reduction in transmitter power.

2 Description of the Utility Function

The two principal components of the utility function for a wireless data system, described in [1], are the signal-to-interference ratio(SIR_i) of each user i, γ_i, and that user's

---

*This work is supported in part by the NSF under Grant No. NCR-07036.

transmitter power P_i. In what follows we consider the uplink of a CDMA data system consisting of one cell and describe utility based on the number of information bits transfered in the lifetime of a user's battery as a function of these two parameters.

Users are distributed at different distances throughout the cell and access the system through a battery powered device. All users share the same bandwidth of W Hertz and encoded data rate R. The SIR_i ratio can be expressed as:

$$\gamma_i = \frac{W}{R} \frac{h_i P_i}{\sum_{j \neq i} h_j P_j + \sigma^2} \tag{1}$$

where the channel gain and power of user i are h_i and P_i, respective, and σ^2 represents the background noise present in the cell. The channel gains, and consequently the location of each user, become significant in determining that user's SIR_i and therefore that user's throughput. The SIR_i value determines the bit error rate experienced by that user via the following equation: $P_{e_i} = Q\left(\sqrt{\gamma_i}\right) \approx 0.5 \cdot e^{-\frac{\gamma_i}{2}}$.

In the present work we incorporate the effects of forward error control into the non-cooperative game by encoding all information into data frames of length L encoded bits. Forward error control is achieved through the use of error correcting codes capable of correcting up to and including t bit errors per frame, and a C bit CRC check included in each frame for error detection. Therefore we can express the probability of correctly receiving an encoded frame, or frame-success-rate (FSR_i) as,

$$FSR_i(\gamma_i, t) = \sum_{j=0}^{t} \binom{L}{j} \left(\frac{e^{-\frac{\gamma_i}{2}}}{2}\right)^{j} \left(1 - \frac{e^{-\frac{\gamma_i}{2}}}{2}\right)^{L-j}. \tag{2}$$

The benefits of forward error control come at the expense of a reduction in the amount of information bits present in each frame. In order to provide error correction it is necessary to insert a number of parity bits into each data frame. A lower bound for the number of parity bits, p, necessary to correct t corrupted bits per frame can be found, as in [2], to be $p = \log_2\left(\sum_{k=0}^{t}\binom{L}{k}\right)$. Including the C CRC check bits we can express the ratio of information bits to encoded bits by $\Gamma(t) = \frac{L-C-p}{L} = \frac{L-C-\log_2\left(\sum_{k=0}^{t}\binom{L}{k}\right)}{L}$. The throughput per battery lifetime for each user is then the rate at which information is correctly transfered through the system multiplied by the lifetime in seconds of the battery:

$$\begin{aligned} Thp_i &= \frac{ER}{P_i} \cdot \Gamma(t) \cdot FSR_i(\gamma, t) \qquad (3)\\ &= \frac{ER}{P_i} \cdot \frac{L - C - \log_2\left(\sum_{k=0}^{t}\binom{L}{k}\right)}{L} \cdot \sum_{j=0}^{t} \binom{L}{j} \left(\frac{e^{-\frac{\gamma_i}{2}}}{2}\right)^{j} \left(1 - \frac{e^{-\frac{\gamma_i}{2}}}{2}\right)^{L-j} \qquad (4) \end{aligned}$$

where E represents the battery energy in Joules. This value corresponds to the total number of useful bits that a user can exchange with the base station during the battery lifetime of their wireless terminal. In the next section we construct a utility function based on this throughput per battery lifetime function.

2.1 Formulation of a Utility Function including Forward Error Control

In [1], the authors detail certain properties of the utility function that yield valid behaviors for the systems considered. In that work however, there was no inclusion of the effects of error correction or error detection. However, the mobile wireless channel is sometimes a hostile environment and error control can greatly reduce the delays caused by the retransmissions of incorrectly received data frames, thus increasing the throughput for the end user. In that vein we construct a utility function that includes both error correction and error detection, and show that it conforms to the five properties in [1].

The measure of satisfaction derived from using a mobile data system clearly is strongly related to the amount of information that one can transfer in the lifetime of their battery. Therefore, it is very convenient to choose the throughput per battery life as the utility function, as this would bring a strong practical and meaningful metric to serve has the definition of utility,

$$U_i = Thp_i. \tag{5}$$

Below, we present the five properties of utility functions described in [1]. We will show that the above choice of utility function is in violation of the first property and will have to be appropriately amended.

The first property that we examine is the limit of user utility as the power tends to zero. We would like there to be no utility derived when the user transmits no power, and this leads to the first property.

Property 1: The utility function, U_i should tend towards zero as the transmitter power tends to zero:

$$\lim_{P_i \to 0} U_i = 0 \tag{6}$$

The choice of throughput per battery lifetime as the utility function violates Property 1, in that it is not forced to zero as the transmitter power is forced to zero. As the transmitter power approaches zero, the resulting $P_{e_i} = \frac{e^{-\frac{\gamma_i}{2}}}{2}$ is forced to 1/2 and the resulting expression becomes:

$$\begin{aligned} \lim_{P_i \to 0} U_i &= \lim_{P_i \to 0} \frac{ER}{P_i} \cdot \frac{L - C - \log_2\left(\sum_{k=0}^{t} \binom{L}{k}\right)}{L} \cdot \sum_{j=0}^{t} \binom{L}{j} \left(\frac{1}{2}\right)^{j} \left(1 - \frac{1}{2}\right)^{L-j} && (7) \\ &= \infty && (8) \end{aligned}$$

So the utility function goes to infinity as the transmitter power goes to zero due to the nonzero value for the FSR_i at zero γ_i. In other words, when the channel is extremely poor, the worst the receiver can do is to randomly guess at the transmitted bits causing the worst bit error performance that can be achieved to be 1/2. As such there is an extremely small, but non-zero, probability that the receiver will guess the right frame, therefore given enough time the user can get an infinite amount of information through the system at no cost to their battery life. In order to handle this pathological case, it was suggested in [1] that the probability of bit error function be modified to yield a zero result when the SIR is zero. This can be done by dropping the 1/2 term in the expression for probability of bit error. Thus a new function is considered that is no longer the exact throughput per battery life but a modified approximation. The proposed utility function can therefore be expressed as

$$\begin{aligned} U_i(P_i, \gamma_i) &= \frac{ER}{P_i} \cdot \frac{L - C - \log_2\left(\sum_{k=0}^{t} \binom{L}{k}\right)}{L} \cdot \sum_{j=0}^{t} \binom{L}{j} \left(e^{-\frac{\gamma_i}{2}}\right)^{j} \left(1 - e^{-\frac{\gamma_i}{2}}\right)^{L-j} && (9) \\ U_i(P_i, \gamma_i) &= \frac{ER\Gamma(t)}{P_i} \cdot f(\gamma_i, t) && (10) \end{aligned}$$

where we denote the modified FSR_i by the function $f(\gamma_i, t)$ which corresponds to the efficiency function described in [1], but for a specific level of error correction.

This function still has practical resonance for a wireless communications system, in that it serves as a underestimate of the number of information bits that can be transmitted during the lifetime of a terminal's battery. It is also an unbiased modification because the new function applies equally well to all users of the system regardless of distance, power or SIR. We feel that, though this function does not have exact practical meaning, valuable insights can be derived from its approximate physical significance.

Property 2: The utility function, U_i should tend towards zero as the transmitter power tends to infinity:

$$\lim_{P_i \to \infty} U_i = 0 \tag{11}$$

Similar to Property 1, Property 2 states that the utility function should be forced to zero as the transmitter power tends to infinity. This is because an infinite transmitter power will burn up the user's battery in zero time, and no information will be sent. We can show that this property holds for our choice of utility function by expanded the efficiency function into its constituent terms and taking the limit.

$$\lim_{P_i \to \infty} \sum_{j=0}^{t} \binom{L}{j} \left(e^{-\frac{\gamma_i}{2}}\right)^j \left(1 - e^{-\frac{\gamma_i}{2}}\right)^{L-j} = \lim_{P_i \to \infty} \left(1 - e^{-\frac{\gamma_i}{2}}\right)^L = 1 \tag{12}$$

$$\tag{13}$$

So the numerator of the utility function approaches a value of $ER\Gamma(t)$, while the denominator goes to infinity as $P_i \to \infty$. Therefore the utility function itself approaches zero, as the transmitter power tends to infinity and the property is satisfied.

Property 3: The utility function, U_i is a monotonically decreasing function of the user's transmitter power, for a fixed γ_i.

$$\frac{\partial U_i(P_i, \gamma_i)}{\partial P_i} < 0, \quad \forall \gamma_i, P_i > 0 \tag{14}$$

Now we are investigating the behavior of the utility as a function of the users transmitted power for a fixed value of γ_i. As the power of a user increases, so does their battery consumption but at no experienced benefit of an increased efficiency function because the SIR value is held constant. Therefore the utility should be a monotonic decreasing function of P_i for a fixed level of interference.

Our utility function is now shown to be consistent with the above property. For a fixed level of interference the partial derivative with respect to the user's transmitter power is:

$$\frac{\partial U_i(P_i, \gamma_i)}{\partial P_i} = -\frac{ER\Gamma(t)}{P_i^2} \cdot f(\gamma_i, t) \tag{15}$$

which is always negative owing to the strict positivity of E, R, Γ and $f(\gamma_i, t)$.

Property 4: The utility function, U_i is a monotonically increasing function of the users SIR_i, γ_i, for a fixed transmitter power.

$$\frac{\partial U_i(P_i, \gamma_i)}{\partial \gamma_i} > 0, \quad \forall \gamma_i, P_i > 0 \tag{16}$$

The CDMA wireless communication system under review here requires that all users access the system through the same over-the-air channel. Thus each additional user serves as an interferer for every other user and degrades the level of satisfaction derived for all other users. The measure of this interference for a particular user i is recorded in the SIR_i, γ_i. Thus the utility should be strongly related to the SIR and as a user's SIR is increased the utility should also increase for the same transmitter power. Our utility function can be found to satisfy this property by showing that the derivative of the efficiency function with respect to γ_i is always positive.

$$\frac{d}{d\gamma_i} f(\gamma_i, t) = \frac{d}{d\gamma_i}\left\{\left(1 - e^{-\frac{\gamma_i}{2}}\right)^L + \ldots + \binom{L}{t}\left(e^{-\frac{\gamma_i}{2}}\right)^t\left(1 - e^{-\frac{\gamma_i}{2}}\right)^{L-t}\right\} \tag{17}$$

$$= \frac{(L-t)\binom{L}{t}}{2}\left(e^{-\frac{\gamma_i}{2}}\right)^{t+1}\left(1 - e^{-\frac{\gamma_i}{2}}\right)^{L-t-1} > 0 \tag{18}$$

This is due to the collapsing series that results from the term by term differentiation. This result can also be obtained by thinking of the efficiency function, $f(\gamma_i, t)$, as a binomial CDF with parameter $e^{-\frac{\gamma_i}{2}}$ and noting that a binomial CDF is monotonically decreasing in its parameter, and as such, is monotonically increasing in γ_i.

Property 5: The utility function, U_i obeys the law of diminishing marginal utility for high values of γ_i.

$$\lim_{\gamma_i \to \infty} \frac{\partial U_i(P_i, \gamma_i)}{\partial \gamma_i} = 0, \quad P_i > 0 \tag{19}$$

The utility should saturate for very high values of γ_i, this is due to the fact that at extremely high values of the γ_i, the channel is nearly perfect and that the added gain for an incrementally better channel is negligible. By taking the limit of (18) as $\gamma_i \to \infty$, we can see that U_i clearly satisfies this property.

3 Distributed Power Control Based on Utility

In this section we discuss the distributed power control scheme. Here we consider the assignment of transmitter powers to K users that are dispersed throughout a cell. The objective of this scheme is to maximize each users utility given the powers of the other users. So the goal can be expressed as a non-cooperative game.

$$\max_{P_i} U_i(P_1, P_2, \ldots, P_i, \ldots, P_K) \quad P_i \in S_i, \quad \forall i = 1, = \ldots, K \tag{20}$$

where S_i is the strictly positive space of all possible values of P_i.

A necessary condition for the equilibrium achieved by such a game is

$$\frac{dU_i}{dP_i} = 0 \quad i = 1, \ldots, K \tag{21}$$

which reduces to the following

$$f(\gamma^*, t) = \gamma^* \cdot f'(\gamma^*, t). \tag{22}$$

Note that the we have dropped the subscript purposely to point to the fact the the first order condition is the same for each user. The above must hold for all K users, which implies that the SIR_i of all the users are equal. Therefore we must satisfy the K linear equations of the form

$$\frac{\frac{W}{R} h_i P_i^*}{\sum_{j \neq i} h_j P_j^* + \sigma} = \gamma^* \quad \forall i \tag{23}$$

where γ^* is the equilibrium operating SIR, to have a feasible power vector $\mathbf{P}^*$. In [1], it is shown that a feasible power vector exists for the single cell case if the following condition is satisfied

$$N \leq \frac{W}{R\gamma^*} + 1. \tag{24}$$

Thus the system capacity limit for a single cell system is determined by the equilibrium solution to the non-cooperative game, γ^*. We will see that the introduction of error control reduces γ^*, and allows a greater number of users into the system. Feasibility for multiple cell systems is presented in [1].

3.1 Existence and Uniqueness of the Equilibrium

We can define the equilibrium solution obtained from the non-cooperative game as a Nash equilibrium, where a Nash equilibrium is a power vector $\mathbf{P}^*$ such that no user can improve its utility by a unilateral change in its power. From [3], Debreau's theorem states that a Nash equilibrium exists for a non-cooperative game if each of the user's utilities are quasi-concave in their own powers.

Theorem 1 *There exists a unique Nash equilibrium to the non-cooperative game in equation (20)*

The existence of a Nash equilibrium follows from the fact that the utility function given in equation (10) is quasi-concave [4] and the use of Debreu's theorem [3]. Uniqueness of the solution follows from the fact that there can only be one solution to equation (22) for the particular U_i chosen.

3.2 Properties of the Equilibrium Solution

3.2.1 Equal Received Powers at the Base Station

It can be shown, as in [1], that the solution to the non-cooperative power control game with forward error correction also achieves equal received powers at the base station.

3.2.2 Pareto Efficiency

A classical measure of the efficiency of an equilibrium solution is Pareto efficiency [5], which states that a solution is Pareto efficient if no user can attain a higher utility without at least one other user suffering a lower utility. Mathematically we can say that a solution $\mathbf{P}^*$ is Pareto efficient if there exists no vector $\mathbf{P}'$ such that $U_{i'} > U_{i^*}$ and $U_{k'} \geq U_{k*}, \forall k = 1...K, j \neq k$, where K is the total number of users in the system.

The solution obtained by the non-cooperative game based on forward error control can be shown to be Pareto inefficient and an across the board decrease in transmitter power of all users leads to higher utilities for each user.

Theorem 2 *There exists some* $0 < \alpha < 1$ *such that for all* i, $U_i(\alpha\mathbf{P}^*, \gamma(\alpha)) > U_i(\mathbf{P}^*, \gamma^*)$, *where* $\gamma(\alpha) = \frac{W}{R}\frac{h_i P_i^*}{\alpha\sum_{j\neq i} h_j P^* j + \sigma^2}$, *and is the SIR associated with the power vector* $\alpha\mathbf{P}^*$

This can be proven by following similar arguments as in [1].

4 Distributed Power Control based on Pricing

As the previous section stated, the equilibrium solution is not Pareto efficient and that higher utilities can be achieved for all users if they all reduce their transmitter power. In this section we describe a pricing function, F_i, applied to the non-cooperative game that encourages this reduction in transmitter powers.

Since the CDMA channel is a shared medium, each user causes interference to every other user. We introduce, $C_{ij} = -\frac{\partial U_i}{\partial P_j} P_j$ as the measure of harm that user j causes to user i. Therefore $C_j = \sum_{i \neq j} C_{ij}$ represents the total harm caused to the system by user j. We would like to develop a pricing function such that users causing greater harm incur greater prices.

Theorem 3 *If the path gains of the K users are in the order $h_1 > h_2 > \ldots > h_k$, then it can be shown that the following is true:*

$$P_1^* < P_2 * < \ldots < P_k^* \tag{25}$$

$$U_1^* > U_2^* > \ldots > U_K^* \tag{26}$$

$$C_1^* < C_2^* < \ldots < C_K^* \tag{27}$$

The proof follows from arguements similiar to that in [1]. The pricing function should then be an increasing function of the user's transmitter power. In this work we consider the same pricing function as that in [1], namely a linear function of the transmitter power.

$$F_i = m \cdot R \cdot P_i. \tag{28}$$

Here m is a positive constant and the pricing function has units of bits, which makes it compatible with the utility function. A user's new goal is then to maximize the difference between utility and price, which can be stated as below;

$$\max_{P_i} U_i - F_i \quad P_i \in S_i, \quad \forall i = 1, \ldots, K \tag{29}$$

This non-cooperative game proceeds as before but with the new objective of utility minus price. We show that a pricing function encourages the use of lower transmit powers and thus can lead to Pareto improvements over the non-cooperative game based solely on utility.

4.1 Existence and Uniqueness of an Equilibrium Solution

It is difficult to say anything analytically about the existence and uniqueness of the solution of the power control game with pricing and forward error control. If the utility-price difference objective function is quasi-concave we can argue as above and show that there exists a Nash equilibrium, but showing the uniqueness of such a solution may be analytically difficult. We do, however, present numerical results that show that an equilibrium does exist and that it does provide Pareto improvements with respect to the the utility only power control game.

5 Numerical Results

In this section we present the numerical results of the non-cooperative game with and without pricing. The parameters of the system under review are listed in Table 1. The non-cooperative power control game was played for a cell containing users at random distances from the base station. The channel gains of the users was determined by the Hata propagation model [6] for the PCS band. The resulting utilities and powers for varying levels of error correction without pricing are shown in Figures 1 and 2 respectively. The improvements due to increasing error correction, in terms of increased user utility are dramatic over the case where no errors are allowed to be corrected.

The introduction of forward error control can be seen to yield Pareto improvements by increasing the utility achieved for all users. These Pareto improvements in the equilibrium solutions of the non-cooperative power control game are realized due to the reduction in the optimal operating SIR target which allows all the users to reduce their transmitter powers in concert. As a consequence of this reduction, the maximum population that a system can sustain and still remain feasible increases as per equation (24). This maximum population and the achieved equilibrium SIR is shown in Figure 3 versus the number correctable errors per frame.

Pricing can be used to achieve a more Pareto efficient equilibrium solution for a non-cooperative power control game for a fixed level of error correction. Particularly pricing can provide sizable Pareto improvements to systems that are operating at or near system capacity. In Figures 4 and 5 we show the effect of pricing, on a full system for various levels of error correction. That is, for a particular value of t the maximum feasible population was allowed into the system and the utilities for the non-cooperative games with and without pricing are plotted versus the user distance from the base station. The graph shows that the utility for all users is higher when a pricing function is introduced.

6 Conclusion

In this work we extended the results of [1] to include the effects of forward error control in the form of error detection and error correction. A utility function based on throughput per battery lifetime including forward error control was developed and shown to yield a unique solution to the non-cooperative power control game. It was shown that error control can yield significant utility increases for all users at a lower transmitter power. Pareto improvements were also seen with increasing forward error correction as well as an increase in the maximum feasible system capacity for the non-cooperative game. A pricing function was also developed and shown to provide further Pareto improvements for systems operating near full system capacity. The combined effect of forward error control and pricing applied to the non-cooperative power control game for wireless data networks was shown to yield favorable utility solutions at reduced transmitter powers.

References

[1] V. Shah, N. Mandayam, and D. Goodman. Power control for wireless data based on utility and pricing. In *Personal, Indoor and Mobile Radio Communications (PIMRC '98)*, Boston, MA, 1998.

[2] J. G. Proakis. *Digital Communication*. McGraw-Hill, New York, NY, 1995.

[3] G. Debreu. A social equilibrium existence theorem. In *Proceedings of the National Academy of Sciences*, volume 38, pages 886–893, 1952.

[4] A. Roberts and D. Varberg. *Convex Functions.* Academic Press, 1973.

[5] H. Varian. *Intermediate Microeconomics, A Modern Approach.* W. W. Norton and Co., New York, NY, 1996.

[6] T. Rappaport. *Wireless Communications Principles and Practice.* Prentice-Hall, New York, NY, 1996.

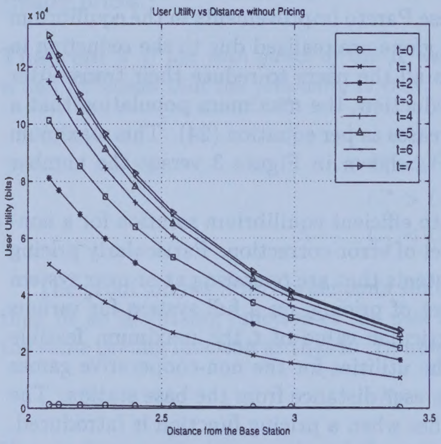


Figure 1: User Utility versus Distance

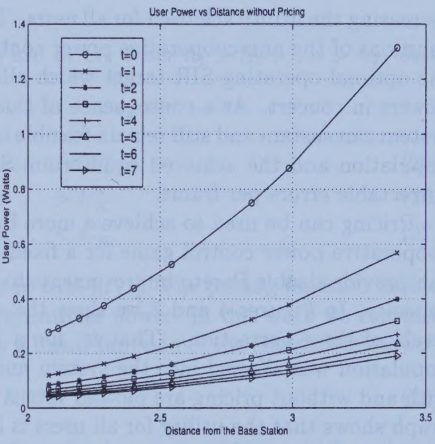


Figure 2: User Power versus Distance

Name	Description	Value
W	Bandwidth of the system	1 MHz
E	Battery energy in Joules	2 Joules
f_c	Carrier frequency	1900 Mhz
R	Encoded Data rate of Users	10 kbps
σ^2	Background noise	5 E-15
h_b	Base Station height	100 meters
h_m	Mobile height	3 meters
L	Encoded data frame length	80 bits
C	Length of CRC code	16 bits

Table 1: System Parameters

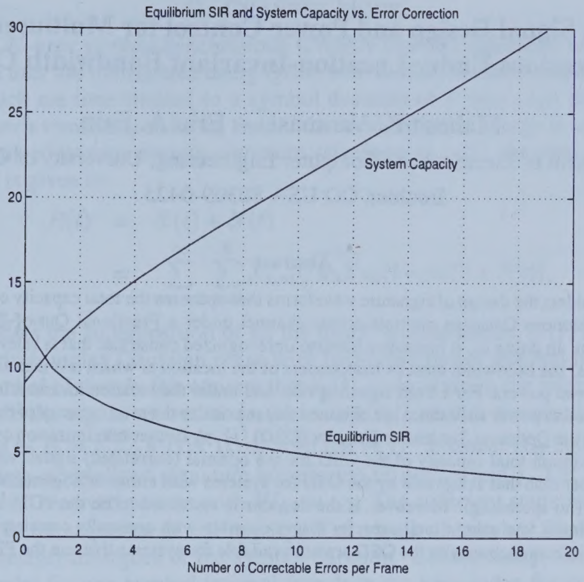


Figure 3: User SIR and System Capacity versus Number of Correctable Errors

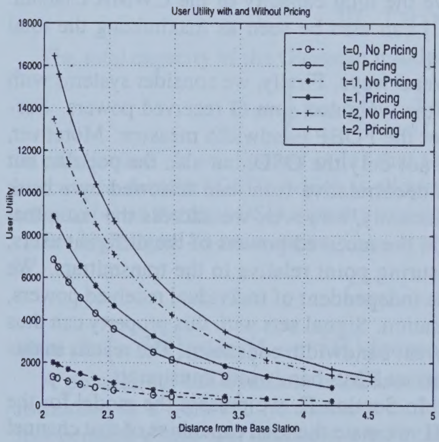


Figure 4: User Utility With and Without Pricing

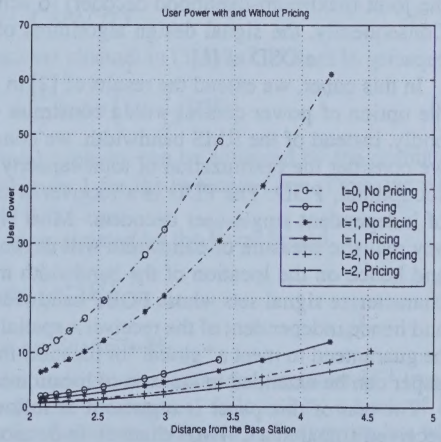


Figure 5: User Power With and Without Pricing

Optimum Signal Design and Power Control for Multiuser Wireless Communications Under Location-Invariant Bandwidth Constraints

Mahesh K. Varanasi and Eric A. Fain
Department of Electrical and Computer Engineering, University of Colorado
Boulder, CO USA 80309-0425

Abstract

This paper considers the design of signature waveforms that optimize the total capacity of multiuser decoders for the QAM-synchronous Gaussian multiple-access channel under a Fractional Out-of-Band Power (FOBP) bandwidth constraint. In doing so, it imposes a hitherto unrecognized constraint that is relevant in wireless communication networks: the bandwidth must be independent of the location at which it is measured, and hence independent of the received powers. For a fixed signaling rate, and under the location invariant bandwidth constraint, signature waveforms and power allocations are obtained that maximize the total capacity of the Post-Decorrelative Decoder (PDD) and the Optimum Successive Decoder (OSD) [1] . A further maximization over the signaling rate reveals that the maximum total capacity of the OSD for the optimal (correlated) signals and signaling rate can be significantly higher than that achievable by the OSD for systems with either orthogonal (maximum spreading) or identical signals (no spreading). Moreover, if the decoder is restricted to be the PDD, orthogonal signaling is nearly optimal. Hence, the gain in total capacity that is possible with optimally correlated signals (relative to orthogonal signals) in conjunction with the OSD are not available for systems that use the PDD.

I. INTRODUCTION

Signature waveforms that maximize the total capacity of the PAM-synchronous, Gaussian, RMS bandwidth constrained, Correlated Waveform Multiple-Access (CWMA) channel were obtained in [2]. It is possible to achieve the total capacity of the CWMA channel by using the OSD obtained in [1] with the implication that it is sufficient to use single-user encoders at the transmitters and the OSD at the receiver (with a much reduced complexity relative to the joint maximum-likelihood decoder) to achieve the total capacity of the CWMA channel. Consequently, the signal design algorithms of [2] can also be seen as maximizing the total capacity of the OSD of [1].

In this paper, we extend the results of [2] in several ways. Firstly, we consider systems with the option of power control with a constraint on the weighted sum of received powers. Secondly, instead of the RMS bandwidth, we consider the FOBP bandwidth measure. Moreover, we consider the optimization of total capacity of not only the OSD, but also the popular, but suboptimal, PDD. The PDD is a receiver with a decorrelative front-end followed by a bank of independent single-user decoders. Most significantly, however, we address the issue that any realistic measure of bandwidth will depend on the received powers of the different users, and hence on the location of the bandwidth measuring point relative to the transmitters. We characterize signal sets whose FOBP bandwidth is independent of individual received powers, and hence independent of the receiver's spatial location. Signal sets with this property can thus be guaranteed to meet a "global" or location-invariant bandwidth constraint. The results in this paper can be extended to the case of location-invariant RMS bandwidth constraints.

The rest of this paper is organized as follows. In Section II, we present the model for the received signal in a CWMA channel. In Section III, we state the total capacities of that channel when the decoder is a PDD or the OSD. In Section IV, we evaluate the FOBP bandwidth measure in the context of the CWMA channel, and introduce the concept of location-invariant bandwidth. In Section V, we present solutions to the signal design and power allocation problems of maximizing the total OSD and PDD capacities under a location-invariant FOBP bandwidth constraint and with a constraint on the weighted sum of received powers. Several numerical examples are given in Section VI. In Section VII, we state the main conclusions of this paper.

This work was supported by NSF Grant NCR-9725778.

II. CHANNEL MODEL

Consider the K-user symbol-synchronous CWMA channel described in [1]. The coded symbols of each user are transmitted using QAM with distinct transmitter pulses or "signature waveforms" which are time-limited to a symbol duration of T secs. Let the received power due to the k^{th} user's transmission be w_k so that the received pulse energy is w_kT. Denoting the k^{th} user's normalized (to unit energy) signature waveform as $s_k(t)$, the total received complex baseband signal is given by

$$\begin{aligned} R(t) &= X(t)+N(t) \qquad (1) \\ &= \sum_{m=-\infty}^{\infty}\sum_{k=1}^{K} X_{k,m}\sqrt{w_kT}\,s_k(t-mT)+N(t), \qquad (2) \end{aligned}$$

where $X(t)$ denotes the superposition of all the users' transmissions, $N(t)$ is zero-mean, circularly symmetric complex Gaussian noise with two-sided power spectral height σ^2. $X_{k,m}$ is the m^{th} code symbol of user k. We assume that $s_k(t)$ is non-zero in the interval $[-T/2, T/2]$.

Let $\mathrm{s}(t)$ denote the signal set vector $[s_1(t) \cdots s_K(t)]^{\mathrm{T}}$ so that the normalized correlation matrix is defined as $\mathrm{R} = \int_{-T/2}^{T/2} \mathrm{s}^*(t)\mathrm{s}^{\mathrm{T}}(t)\,dt$ with the $(i,j)^{th}$ element R_{ij} being the correlation between the energy-normalized signature waveforms of users i and j. Let W be a diagonal matrix of the users' received powers with $\mathrm{W}_{kk} = w_k$. The sufficient statistics for the model in (1) are obtained as the sequence of vector outputs of a bank of matched filters matched to $\mathrm{s}(t)$ and sampled at integer multiples of the symbol duration. Consequently, the channel admits a discrete-time model for one symbol interval (we drop the time-index for convenience) given by

$$\underline{Y} = \mathrm{H}\underline{X} + \underline{N}, \qquad (3)$$

where $\mathrm{H} = T\mathrm{W}^{1/2}\mathrm{R}\mathrm{W}^{1/2}$ is the energy-weighted correlation matrix, $\underline{X}$ is a vector of the users' transmitted code symbols, and $\underline{N}$ is a circularly-symmetric complex Gaussian noise vector with covariance matrix $\sigma^2\mathrm{H}$. Note that the received signal energy of the k^{th} signal w_kT is inversely proportional to the symbol rate.

III. CHANNEL CAPACITY

The total capacity of the Gaussian multiple-access channel in (3) is (cf. [1])

$$C = \frac{1}{T}\ln\det\left(\mathrm{I}+\frac{1}{\sigma^2}\mathrm{H}\right) = \frac{1}{T}\ln\det\left(\mathrm{I}+\frac{T}{\sigma^2}\mathrm{R}\mathrm{W}\right) \qquad (4)$$

in nats/second. It was proved in [1] that this total capacity can be achieved by the optimum successive decoder (OSD), which is shown in Figure 1 for the two-user case. Thus, we will refer to (4) as the total OSD capacity and denote it as C_{OSD}.

For high signal-to-noise ratios, C_{OSD} is well approximated by the lower bound

$$C_{\mathrm{OSD}} \geq \frac{1}{T}\ln\det\left(\sigma^{-2}\mathrm{H}\right), \qquad (5)$$

since $\det\left(\sigma^{-2}\mathrm{H}\right) \leq \det\left(\mathrm{I}+\sigma^{-2}\mathrm{H}\right)$ for positive semidefinite H.

A popular but sub-optimum multiuser decoder consists of a linear decorrelative H^{-1} transformation as a front-end (to tune-out multiple-access interference) followed by a bank of K independent single-user decoders, one for each user as shown in Figure 2. For brevity, we will refer to this decoder as the post-decorrelative decoder (PDD). This structure effectively creates K single-user channels with noise variance $\sigma^2\mathrm{R}^{-1}_{kk}$. The total capacity of these K channels is therefore given by

$$C_{\mathrm{PDD}} = \frac{1}{T}\sum_{k=1}^{K}\ln\left(1+\frac{w_kT}{\sigma^2}[\mathrm{R}^{-1}_{kk}]^{-1}\right), \qquad (6)$$

which we will call the total PDD capacity of the channel.

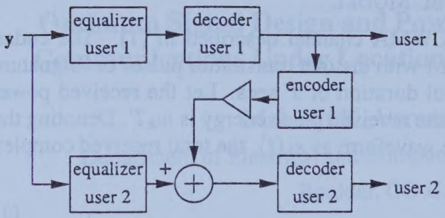


Fig. 1. The optimum successive decoder (OSD) for a two-user channel.

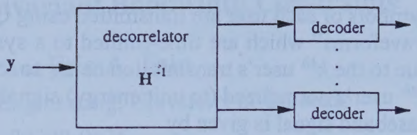


Fig. 2. The post-decorrelative decoder (PDD).

IV. FRACTIONAL OUT-OF-BAND POWER

Bandwidth can be defined in many different ways for time-limited signals (e.g. [3]), some of which are more analytically tractable than others. In this paper, we will focus on the fractional out-of-band power as our bandwidth measure.

Definition 1: A baseband wide-sense stationary (WSS) random process $U(t)$ with power spectral density $S_U(f)$ has a fractional out-of-band power (FOBP) $\epsilon(U)$ for the band $[-B,B]$ given by

$$\epsilon(U) = \frac{\int_{|f|>B} S_U(f)\,df}{\int_{-\infty}^{\infty} S_U(f)\,df}\,. \tag{7}$$

The definition of FOBP can be extended to processes which are wide-sense cyclostationary with period T with $S_U(f)$ denoting the power spectral density of the associated WSS process $U(t-\Theta)$ where Θ is a uniformly distributed random variable over $[0,T]$. Moreover, if $u(t)$ is a deterministic, baseband, finite energy function, with energy spectral density $S_U(f)$, then (7) defines its fractional out-of-band energy (FOBE).

A. Signal Space

We will assume that the signature waveforms are restricted to the class $\mathcal{L}^2_{[-T/2,T/2]}$ which consists of all signals whose real and imaginary parts are square-integrable on the interval $[-T/2,T/2]$. There are many possible complete orthonormal bases for this space (for instance, the traditional Fourier bases) so that any function can be written as a linear combination of basis functions $\{\psi_i(t)\}$ as $s(t) = \sum_{i=0}^{\infty} g_i \psi_i(t)$, with convergence in the $\mathcal{L}^2$ norm. It is convenient to number the basis functions according to the non-decreasing order of their fractional out-of-band energies relative to the band $[-B,B]$ so that $\epsilon(\psi_0) \leq \epsilon(\psi_1) \leq \epsilon(\psi_2) \leq \cdots$.

Let us define the infinite-dimensional vector of the orthonormal basis functions $\Psi(t)$ as

$$\Psi(t) = [\psi_0(t) \quad \psi_1(t) \quad \psi_2(t) \quad \cdots \quad]^{\mathrm{T}}, \tag{8}$$

so that $s_k(t)$ can be expressed as $\Psi^{\mathrm{T}}(t)\mathrm{g}_k$ with g_k being an infinite-dimensional complex-valued vector. Since $s_k(t)$ has unit energy, and $\int \Psi(t)\Psi^{\mathrm{T}}(t)dt = \mathrm{I}$, g_k must be normalized so that $\mathrm{g}_k^{\dagger}\mathrm{g}_k = 1$. Further, let $\Psi_i(f)$ be the Fourier transform of $\psi_i(t)$, and define the matrix M so that its i,j^{th} element is

$$\mathrm{M}_{ij} = \int_{|f|>B} \Psi_{i-1}^{*}(f)\Psi_{j-1}(f)\,df, \tag{9}$$

so that $\mathrm{M}_{ii} = \epsilon(\psi_{i-1})$. It is easily seen that the FOBE of $s_k(t)$ is given by $\mathrm{g}_k^{\dagger}\mathrm{M}\mathrm{g}_k/\mathrm{g}_k^{\dagger}\mathrm{g}_k$.

Let us denote the signal set vector $\mathrm{s}(t)$ as

$$\mathrm{s}(t) = \mathrm{G}^{\mathrm{T}}\Psi(t)\,, \tag{10}$$

so that g_k is the k^{th} column of G. We will henceforth refer to G as the signature sequence matrix with g_k denoting the signature sequence of user k.

The power spectral density (cf. [4]) of the wide-sense cyclostationary signal $X(t)$, which is the signal component of the received signal in (1), assuming that each user's data sequence is white, zero-mean, and uncorrelated with the other users' data sequences, is given as

$$S_X(f) = \sum_{k=1}^{K} w_k |S_k(f)|^2, \tag{11}$$

where $S_k(f)$ denotes the Fourier transform of $s_k(t)$. The FOBP of $X(t)$ is then given by

$$\epsilon(X) = \frac{\sum_k w_k \mathrm{g}_k^\dagger \mathrm{M} \mathrm{g}_k}{\sum_k w_k} = \frac{\mathrm{tr}(\mathrm{G}^\dagger \mathrm{M} \mathrm{G} \mathrm{W})}{\mathrm{tr}(\mathrm{W})}. \tag{12}$$

Given T and W, the total OSD and PDD capacities in (4) and (6) depend on the signature sequence matrix G only through the correlation matrix

$$\mathrm{R} = \int_{-T/2}^{T/2} \mathrm{s}^*(t)\mathrm{s}^\mathrm{T}(t)\, dt = \mathrm{G}^\dagger \mathrm{G}, \tag{13}$$

and are independent of the choice of basis functions. However, the FOBP of $X(t)$ depends both on G and the basis functions through M.

B. Location-Invariant Bandwidth

The FOBP of a signal set is dependent on the individual received pulse energies through W as seen in (12). This means that if we impose a FOBP constraint at the receiver, it is not guaranteed to be satisfied at other spatial locations, such as at cell boundaries and in adjacent cells, where the received powers will differ, but where the constraint must also be imposed. It is therefore useful to characterize signal sets that yield an FOBP which is independent of the receiver's location relative to the transmitters.

Definition 2 (Location-Invariant FOBP) A signal set whose FOBP is independent of the received powers W is said to have a location-invariant FOBP.

Proposition 1: The signal set $\mathrm{s}(t) = \mathrm{G}^\mathrm{T}\Psi(t)$ has location-invariant FOBP if and only if $\mathrm{G}^\dagger\mathrm{M}\mathrm{G}$ has equal diagonal elements. In this case, the FOBP is $\mathrm{tr}(\mathrm{G}^\dagger\mathrm{M}\mathrm{G})/K$.

Proof: The FOBP is

$$\frac{\mathrm{tr}(\mathrm{G}^\dagger \mathrm{M} \mathrm{G} \mathrm{W})}{\mathrm{tr}(\mathrm{W})} = \frac{\sum_{i=1}^{K} \mathrm{W}_{ii}[\mathrm{G}^\dagger \mathrm{M} \mathrm{G}]_{ii}}{\sum_{i=1}^{K} \mathrm{W}_{ii}} \tag{14}$$

so we have sufficiency. Now suppose $\mathrm{G}^\dagger\mathrm{M}\mathrm{G}$ does not have equal diagonal elements. There must exist a W, (for example, $\mathrm{W} = \mathrm{diag}(\mathrm{G}^\dagger\mathrm{M}\mathrm{G})$) for which the FOBP bandwidth measure is a function of W. Hence the proposition. ■

The issue of W dependence occurs for the RMS bandwidth constraint treated in [2] as well, and a similar characterization can be obtained for the location-invariant RMS bandwidth.

V. CAPACITY-MAXIMIZING SIGNAL DESIGN WITH POWER CONTROL

The objective is to find signal sets and power allocations that jointly optimize the total OSD or PDD capacities. Each signature signal belongs to the class of time-limited and energy-normalized signals in $\mathcal{L}^2_{[-T/2,T/2]}$ and the K signals as a set are constrained to satisfy a FOBP constraint $\epsilon(X) \leq b_0$ and the location-invariant condition of Proposition 1. There is also a constraint on the weighted sum of received powers. Such a constraint can be expressed as

$$\mathrm{tr}(\mathrm{WA}) = P, \tag{15}$$

where $\mathbf{A}$ is a diagonal matrix of positive weights. In general $\mathbf{A}_{kk}$ can be interpreted as the cost per unit power received at the base station from user k so that $\mathrm{tr}(\mathbf{WA})$ denotes the total cost. The cost of received power from a particular transmitter may depend on several factors such as the distance from that transmitter to the base station as well as on the cost of power generation at that transmitter. When the diagonal elements of $\mathbf{A}$ are chosen to be inversely proportional to the path losses for the K users, then (15) denotes a constraint on the total transmitted power.

The above problems are difficult to solve outright. However, if the total OSD capacity is replaced by its lower bound in (5), which is tight for high SNR, it can be shown that the corresponding optimum signal design and power control problem can be decoupled into problems of finding optimal signature sequences, optimal basis functions, and optimal power allocations and eigenvalues of the correlation matrix of the signature waveforms. With a slight restriction on the class of signal sets, such a decoupling is also possible for the optimization of the total PDD capacity. The rest of this paper obtains solutions to precisely those decoupled problems.

A. Optimal Sequences

Given a set of basis functions, the optimal signature sequences can be shown to minimize the FOBP bandwidth under the location-invariant condition of Proposition 1. To find the optimal sequences, it is convenient to let $\mathbf{M}$ have a spectral decomposition $\mathbf{M} = \mathbf{Y}^{\dagger}\mathbf{\Pi}\mathbf{Y}$, with the eigenvalues in $\mathbf{\Pi}$ arranged in *non-decreasing* order. We have the following proposition:

Proposition 2 (Minimum location-invariant FOBE sequences) Given a K-dimensional diagonal matrix $\mathbf{\Lambda}$ of the eigenvalues of $\mathbf{R}$, with entries arranged in *non-increasing* order, the optimal signature sequence matrix $\hat{\mathbf{G}}$ that satisfies the condition in Proposition 1 and the normalization constraint $\mathrm{diag}\left\{\mathbf{G}^{\dagger}\mathbf{G}\right\} = \mathbf{I}$ and minimizes the location-invariant bandwidth is given by $\hat{\mathbf{G}} = \tilde{\mathbf{Y}}^{\dagger}\mathbf{\Lambda}^{1/2}\mathbf{U}$, where $\mathbf{U}$ is a scaled DFT matrix [5] (or a scaled Hadamard matrix for all K for which Hadamard matrices exist [6]), and $\tilde{\mathbf{Y}}$ is the $(K \times \infty)$ matrix of the first K rows of $\mathbf{Y}$. Furthermore, for this optimal choice of the signature sequences, the FOBP admits the formula

$$\epsilon(X) = K^{-1}\mathrm{tr}\left(\mathbf{\Lambda}\tilde{\mathbf{\Pi}}\right), \tag{16}$$

where $\tilde{\mathbf{\Pi}}$ is the upper-left $K \times K$ diagonal submatrix of $\mathbf{\Pi}$.

Proof: Let $\mathbf{G}$ satisfy the location-invariance property of Proposition 1, so that $\mathbf{G}^{\dagger}\mathbf{M}\mathbf{G}$ has equal diagonals. The signature correlation matrix $\mathbf{R} = \mathbf{G}^{\dagger}\mathbf{G}$ has spectral decomposition $\mathbf{U}^{\dagger}\mathbf{\Lambda}\mathbf{U}$, so that in general, $\mathbf{G} = \mathbf{X}\mathbf{\Lambda}^{1/2}\mathbf{U}$ for some $\infty \times K$ matrix $\mathbf{X}$ which satisfies $\mathbf{X}^{\dagger}\mathbf{X} = \mathbf{I}_K$ (identity matrix). The problem is to minimize the FOBP, $K^{-1}\mathrm{tr}(\mathbf{G}^{\dagger}\mathbf{Y}^{\dagger}\mathbf{\Pi}\mathbf{Y}\mathbf{G})$, over $\mathbf{X}$ and $\mathbf{U}$ under the location-invariance and normalization constraints. Now,

$$\begin{aligned} \mathrm{tr}(\mathbf{G}^{\dagger}\mathbf{Y}^{\dagger}\mathbf{\Pi}\mathbf{Y}\mathbf{G}) &= \mathrm{tr}\left[\mathbf{U}^{\dagger}\mathbf{\Lambda}^{1/2}\mathbf{X}^{\dagger}\mathbf{Y}^{\dagger}\mathbf{\Pi}\mathbf{Y}\mathbf{X}\mathbf{\Lambda}^{1/2}\mathbf{U}\right] & (17)\\ &= \mathrm{tr}(\mathbf{\Lambda}\mathbf{X}^{\dagger}\mathbf{Y}^{\dagger}\mathbf{\Pi}\mathbf{Y}\mathbf{X}) = \mathrm{tr}(\mathbf{\Lambda}\mathbf{Q}) & (18) \end{aligned}$$

where $\mathbf{Q} = \mathbf{X}^{\dagger}\mathbf{Y}^{\dagger}\mathbf{\Pi}\mathbf{Y}\mathbf{X}$, $(K \times K)$. Using a trace inequality [7, (H.1.h)],

$$\mathrm{tr}(\mathbf{\Lambda}\mathbf{Q}) \geq \sum_{i=1}^{K} \phi_{K-i+1}(\mathbf{\Lambda})\phi_i(\mathbf{Q}), \tag{19}$$

(with $\phi_i(\mathbf{Q})$ denoting the i^{th} smallest eigenvalue of $\mathbf{Q}$). By the Poincaré separation theorem [8, (4.3.16)], $\phi_i(\mathbf{Q}) \geq \phi_i(\mathbf{\Pi})$, so

$$\mathrm{tr}(\mathbf{\Lambda}\mathbf{Q}) \geq \sum_{i=1}^{K} \mathbf{\Lambda}_{ii}\mathbf{\Pi}_{ii}, \tag{20}$$

and equality is achieved by letting $\mathbf{X}^{\dagger} = [\mathbf{I}_K\, 0]\mathbf{Y} = \tilde{\mathbf{Y}}$. Fortunately, for this minimum-bandwidth choice of $\mathbf{X}$, the location-invariance and normalization constraints can always be satisfied by choosing the unitary matrix U to be a scaled DFT (or Hadamard when K is a Hadamard dimension) matrix, since such a matrix has the property that $|\mathrm{U}_{ij}|^2 = K^{-1}$ for all i, j, so that

$$\epsilon(X) = [\mathrm{G}^\dagger \Pi \mathrm{G}]_{kk} = [\mathrm{U}^\dagger \Lambda \Pi \mathrm{U}]_{kk} = \sum_{i=1}^{K} |\mathrm{U}_{ik}|^2 \Lambda_{ii} \Pi_{ii} = \mathrm{tr}(\Lambda \Pi)/K, \tag{21}$$

and similarly, $\mathrm{diag}\left\{\mathrm{G}^\dagger \mathrm{G}\right\} = \mathrm{I}$. ■

In the remainder of this paper, we use the location-invariant bandwidth constraint to imply that U is a DFT (or Hadamard) matrix.

B. Optimal Basis Functions

Let us define the prolate-spheroidal wave functions [9]: for a given time T and bandwidth B, the prolate-spheroidal wave functions $\{\phi_i(t)\}_{i=0}^{\infty}$ are a complete orthonormal basis for the space of all square-integrable functions band-limited to $[-B, B]$. The time-truncation of $\phi_i(t)$ to $[-T/2, T/2]$ has a fraction χ_i of its total energy, and furthermore,

$$\int_{-\infty}^{\infty} \phi_i(t)\phi_j(t)\,dt = \delta_{ij}, \qquad \int_{-T/2}^{T/2} \phi_i(t)\phi_j(t)\,dt = \chi_i \delta_{ij}, \tag{22}$$

where δ_{ij} is the Kronecker delta, so that the set of all the *time-truncated* $\phi_i(t)$'s are a complete orthonormal basis for the space of time-limited real square-integrable functions. Moreover, the band-limited version (to $[-B, B]$) of the time-limited function $\phi_i(t)$ has energy χ_i^2. Indexing these time-limited functions in decreasing order of their energies so that

$$\chi_0 > \chi_1 > \chi_2 > \cdots. \tag{23}$$

Evidently, the FOBE of $\phi_i(t)$ truncated to $[-T/2, T/2]$ is $1 - \chi_i$, and we will define $\hat{\phi}_i(t) = \phi_i(t)/\sqrt{\chi_i}$ on $[-T/2, T/2]$. The computation of the χ_i's and the prolate-spheroidal wave functions is described in [10] and elsewhere.

Proposition 3: Given Λ, the diagonal matrix of the eigenvalues of R, with entries arranged in *non-increasing* order, the optimum signal set (with minimum FOBP) is given as $\mathrm{s}^T(t) = \hat{\Phi}^T(t)\Lambda^{1/2}\mathrm{U}$, with U chosen for location-invariance and where $\hat{\Phi}^T(t)$ contains the first K (smallest FOBE) prolate-spheroidal wave functions, i.e., $\hat{\Phi}^T(t) = [\hat{\phi}_0(t), \hat{\phi}_1(t), \cdots, \hat{\phi}_{K-1}(t)]^T$.

Proof: Given any arbitrary orthonormal basis $\Psi^T(t)$, consider the rotated orthonormal basis $\Phi^T(t) = \Psi^T(t)\mathrm{Y}^\dagger$ where Y diagonalizes M. Consequently, $\int_{|f|>B} \Phi^*(f)\Phi^T(f)df = \Pi$ so that the rotated basis functions in $\Phi^T(t)$ are arranged in the non-decreasing order of their FOBEs with $\epsilon(\phi_i) = \Pi_{ii}$. Proposition 2 implies that given Λ, the signal set with the optimum signature sequence matrix is of the form $\tilde{\Phi}^T(t)\Lambda^{1/2}\mathrm{U}$, with U chosen for location-invariance, and where the K-dimensional vector $\tilde{\Phi}^T(t)$ is obtained by retaining the first K, smallest FOBE functions in $\Phi^T(t)$. That the optimal $\tilde{\Phi}^T(t)$ that further minimizes the minimum FOBP in (16) is indeed $\hat{\Phi}^T(t)$ follows from the fact that $\{\hat{\phi}_i(t)\}_{i=0}^{K-1}$ have the property that for all k, $\hat{\phi}_k(t)$ has the minimum FOBE out of all $\mathcal{L}^2$ functions which are orthogonal to $\{\hat{\phi}_i(t)\}_{i=0}^{k-1}$ [9]. The corresponding diagonal matrix of FOBEs is $\hat{\Pi} = \mathrm{diag}[1-\chi_0, 1-\chi_1, \cdots, 1-\chi_{K-1}]$. The fact that the matrix $\int_{|f|>B} \hat{\Phi}^*(f)\hat{\Phi}^T(f)df = \hat{\Pi}$ (and hence diagonal) is the frequency dual of (22). ■

In the remainder of this paper, we will denote $\hat{\Pi}$ as Π. Note that when K is a Hadamard dimension, the optimum signals can be chosen to be real-valued.

C. Optimization of total OSD capacity

Proposition 4: The problem of finding the signature waveforms and powers to maximize total OSD capacity over signal sets that satisfy the FOBP bandwidth constraint $\epsilon(X) \leq b_0$ and the location-invariance condition of Proposition 1 and over received powers that satisfy the weighted sum power constraint can now be stated in terms of Λ and $\mathbf{W}$ as:

$$\hat{\Lambda}_{\mathrm{OSD}}, \hat{\mathbf{W}}_{\mathrm{OSD}} = \arg \max_{\Lambda, \mathbf{W}} \det(\mathbf{I} + \tfrac{T}{\sigma^2}\mathbf{U}^{\dagger}\Lambda\mathbf{U}\mathbf{W}) \tag{24}$$

$$\text{s.t. } \mathrm{tr}(\Lambda\Pi) \leq Kb_0,\ \mathrm{tr}(\Lambda) = K,\ \mathrm{tr}(\mathbf{W}\Lambda) = P,\ \Lambda \geq 0,\ \Lambda_{i+1,i+1} \geq \Lambda_{ii}\ i = 1, \cdots, K-1$$

where $\mathbf{U}$ is chosen for location-invariance. The vector of optimum signal waveforms is then given by $\mathbf{s}^{\mathrm{T}}(t) = \hat{\Psi}^{\mathrm{T}}(t)\hat{\Lambda}_{\mathrm{OSD}}^{1/2}\mathbf{U}$.

In general, the optimization in (24) is a difficult problem which depends on the choice of $\mathbf{U}$. Thus, we consider the optimization of the lower bound on total OSD capacity. Since $\det(\mathbf{H}) = T\det(\mathbf{R})\det(\mathbf{W})$, we have two independent optimizations over Λ and $\mathbf{W}$.

Proposition 5: The problem of finding the signature waveforms to maximize the lower bound on total capacity under a location-invariant bandwidth constraint can be posed entirely in terms of Λ and $\mathbf{W}$ as

$$\hat{\Lambda}_{\mathrm{lb}} = \arg \max_{\Lambda} \det(\Lambda) \tag{25}$$

$$\text{s.t. } \mathrm{tr}(\Lambda\Pi) \leq Kb_0,\ \mathrm{tr}(\Lambda) = K,\ \Lambda > 0,\ \Lambda_{i+1,i+1} \geq \Lambda_{ii}\ i = 1, 2, \cdots, K-1$$

and

$$\hat{\mathbf{W}}_{\mathrm{lb}} = \arg \max_{\mathbf{W}} \det(\mathbf{W}) \tag{26}$$

$$\text{s.t. } \mathrm{tr}(\mathbf{W}\Lambda) = P,\ \mathbf{W} > 0.$$

The vector of optimum signal waveforms is then given by $\mathbf{s}^{\mathrm{T}}(t) = \hat{\Psi}^{\mathrm{T}}(t)\hat{\Lambda}_{\mathrm{lb}}^{1/2}\mathbf{U}$, where $\mathbf{U}$ is a scaled DFT (Hadamard) matrix.
We denote the correlation matrix of this signal set as $\mathbf{H}_{\mathrm{lb}} = T\hat{\mathbf{W}}_{\mathrm{lb}}^{1/2}\mathbf{U}^{\dagger}\hat{\Lambda}_{\mathrm{lb}}\mathbf{U}\hat{\mathbf{W}}_{\mathrm{lb}}^{1/2}$. The problem in (25) can be solved using Lagrange multipliers as in [2]. It requires the solution of the nonlinear equation

$$\sum_{k=1}^{K} \frac{1}{1 - \mu(\Pi_{kk} - b_0)} = K \tag{27}$$

for the parameter μ with $\mu \leq (\Pi_{KK} - b_0)^{-1}$. The eigenvalues are then given by $\Lambda_{kk} = \lambda_k = [1 - \mu(\Pi_{kk} - b_0)]^{-1}$. Note that the valid range for b_0 is

$$\Pi_{11} < b_0 < \mathrm{tr}(\Pi)/K. \tag{28}$$

The optimization in (26) is easily solved in closed form as

$$\hat{\mathbf{W}}_{\mathrm{lb}} = \frac{P}{K}\Lambda^{-1}. \tag{29}$$

D. Optimization of total PDD capacity

Under a location-invariant bandwidth constraint, for any diagonal matrix of eigenvalues Λ, from Proposition 2, the correlation matrix is given by $\mathbf{R} = \mathbf{U}^{\dagger}\Lambda\mathbf{U}$, where $\mathbf{U}$ has the property that $|\mathbf{U}_{ij}|^2 = K^{-1}$. Since the inverse is written as $\mathbf{R}^{-1} = \mathbf{U}^{\dagger}\Lambda^{-1}\mathbf{U}$, we have that

$$[\mathbf{R}^{-1}_{kk}]^{-1} = K\left(\sum_{i=1}^{K}\frac{1}{\lambda_i}\right)^{-1} = K/\mathrm{tr}(\Lambda^{-1}) \tag{30}$$

which is the harmonic mean of the eigenvalues ($\lambda_i = \Lambda_{ii}$). Substituting (30) into (6), the PDD capacity can now be written as

$$C_{\text{PDD}} = \frac{1}{T} \ln \det \left(\mathbf{I} + \frac{TK}{\sigma^2 \text{tr}(\Lambda^{-1})} \mathbf{W} \right). \tag{31}$$

Hence, maximizing the PDD capacity for any $\mathbf{W}$ requires the maximization of the harmonic mean of the eigenvalues of $\mathbf{R}$. Thus, we can first find the optimum Λ, then find the optimum $\mathbf{W}$ given Λ.

Proposition 6: The problem of finding signature waveforms and powers that maximize the total PDD capacity under a location-invariant bandwidth constraint and weighted average power constraint can be posed entirely in terms of Λ and $\mathbf{W}$ as

$$\hat{\Lambda}_{\text{PDD}} = \arg \min_{\Lambda} \text{tr}(\Lambda^{-1}) \tag{32}$$

$$\text{s.t. } \text{tr}(\Lambda\Pi) \leq Kb_0,\ \text{tr}(\Lambda) = K,\ \Lambda > 0,\ \Lambda_{i+1,i+1} \geq \Lambda_{ii}\ i = 1, 2, \cdots, K-1$$

and

$$\hat{\mathbf{W}}_{\text{PDD}} = \arg \max_{\mathbf{W}} \det(\mathbf{I} + \tfrac{TK}{\sigma^2} \text{tr}(\hat{\Lambda}_{\text{PDD}}^{-1})^{-1} \mathbf{W}) \tag{33}$$

$$\text{s.t. } \text{tr}(\mathbf{W}\mathbf{A}) = P,\ \mathbf{W} > 0.$$

The vector of optimum signal waveforms is then given by $\mathbf{s}^T(t) = \hat{\Psi}^T(t) \hat{\Lambda}_{\text{PDD}}^{1/2} \mathbf{U}$.

We denote the correlation matrix of this signal set as $\mathbf{H}_{\text{PDD}} = T\hat{\mathbf{W}}_{\text{PDD}}^{1/2} \mathbf{U}^{\dagger} \hat{\Lambda}_{\text{PDD}} \mathbf{U} \hat{\mathbf{W}}_{\text{PDD}}^{1/2}$. The problem in (32) can also be solved using Lagrange multipliers. Ignoring for now the positivity constraints, we have the Lagrangian

$$L(\Lambda, \mu_1, \mu_2) = \sum_{i=1}^{K} \frac{1}{\lambda_i} + \mu_1 \left(\sum_{i=1}^{K} \lambda_i - K \right) + \mu_2 \left(\sum_{i=1}^{K} \lambda_i \Pi_{ii} - Kb_0 \right). \tag{34}$$

Setting the derivative $\frac{\partial}{\partial \lambda_i} L = 0$ yields

$$\lambda_i = (\mu_1 + \mu_2 \Pi_{ii})^{-1/2}. \tag{35}$$

Solving the two equality constraints for the Lagrange multipliers (μ_1, μ_2) results in the two nonlinear equations

$$\sum_{i=1}^{K} \frac{1}{\sqrt{\mu_1 + \mu_2 \Pi_{ii}}} = K \qquad \sum_{i=1}^{K} \frac{\Pi_{ii}}{\sqrt{\mu_1 + \mu_2 \Pi_{ii}}} = Kb_0 \tag{36}$$

which can be solved numerically for with the constraint that $\mu_1 + \mu_2 \Pi_{ii} > 0$ to ensure that $\Lambda > 0$. Again, the valid range for b_0 is $\Pi_{11} < b_0 < \text{tr}(\Pi)/K$.

Solving for the optimal $\mathbf{W}$ is accomplished with the Lagrangian

$$L(\mathbf{W}, \mu) = \sum_{i=1}^{K} \ln \left(1 + \frac{TK}{\sigma^2} \text{tr}(\hat{\Lambda}_{\text{PDD}}^{-1})^{-1} w_i \right) - \mu \left(\sum_{i=1}^{K} w_i a_i - P \right). \tag{37}$$

Taking the derivative and solving for w_i, we have

$$\frac{1}{1 + \frac{TK}{\sigma^2} \text{tr}(\hat{\Lambda}_{\text{PDD}}^{-1})^{-1} w_i} = \mu a_i \qquad \text{and} \qquad w_i = \frac{\sigma^2}{TK} \text{tr}(\hat{\Lambda}_{\text{PDD}}^{-1}) \left[\frac{1}{\mu a_i} - 1 \right] \tag{38}$$

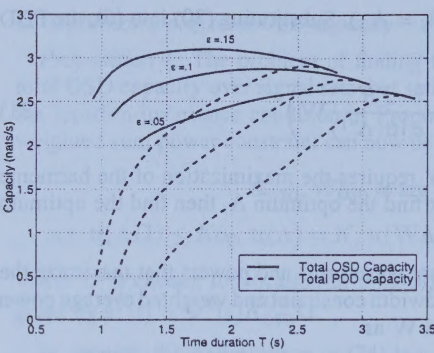


Fig. 3. Maximum total capacity for a four-user example as a function of symbol duration T. Values of FOBP relative to a band B of 4 rad/s are .05, .1, and .15. The average SNR is 3dB.

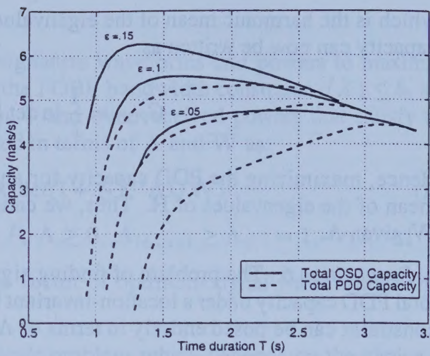


Fig. 4. Maximum total capacity for a four-user example as a function of symbol duration T. Values of FOBP relative to a band B of 4 rad/s are .05, .1, and .15. The average SNR is 10dB.

so

$$\hat{\mathbf{W}}_{\text{PDD}} = \left(\frac{P}{K} + \operatorname{tr}(\mathbf{A})\operatorname{tr}(\hat{\mathbf{\Lambda}}_{\text{PDD}}^{-1})\frac{\sigma^2}{K^2T}\right)\mathbf{A}^{-1} - \operatorname{tr}(\hat{\mathbf{\Lambda}}_{\text{PDD}}^{-1})\frac{\sigma^2}{KT}\mathbf{I} \tag{39}$$

which reduces to (29) for small σ^2.

We remark here as an aside that, since the harmonic mean of a set of positive numbers lower bounds the geometric mean of those numbers, the eigenvalues that optimize the total PDD capacity can also be seen to optimize a lower bound on the total OSD capacity.

VI. NUMERICAL RESULTS

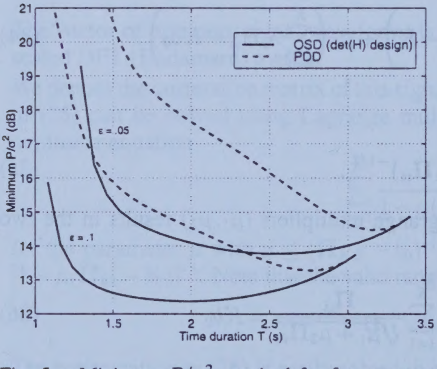


Fig. 5. Minimum P/σ^2 required for four users to achieve a capacity of 4 nats/s, with $\operatorname{tr}(\mathbf{WA}) = P$. Values of FOBP relative to a band B of 4 rad/s are .05 and .1, diag(A)=[1.66, 1.11, .74, .49].

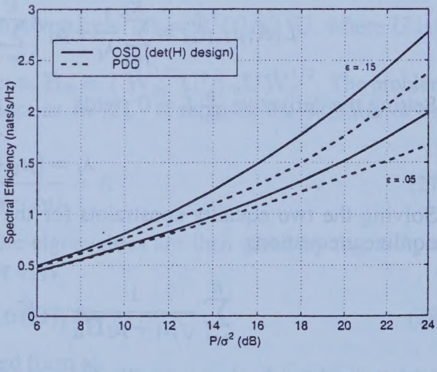


Fig. 6. Maximum spectral efficiency C/B for four users. Values of FOBP relative to a band B of 4 rad/s are .05 and .15, with $\operatorname{tr}(\mathbf{WA}) = P$, $\mathbf{A} = \mathbf{I}$.

We now investigate capacity as a function of the signaling rate. To this end, we consider the band B for which we measure the FOBP to be fixed, and we vary the signal duration T. It is known [9] that for a given B and FOBP ϵ, there is a minimum value of T. For this B and T, we will have $\chi_0 = \epsilon$, and so our bandwidth constraint will require identical waveform signaling ($s_k(t) = \hat{\psi}_0(t)$ for all k). Additionally, there will be a value of T for which the signals can be made orthogonal and still meet the FOBP constraint. In Figures 3 and 4, the total OSD and PDD capacities are plotted for a four-user case with diag(A) = [1.66, 1.11, 0.74, 0.49] for several values of FOBP ($\epsilon = .05,\ 0.1,\ 0.15$) for the corresponding range of T. The ratio of average transmitted power to received noise level ($P/K\sigma^2$) is fixed at 3dB (Fig. 3) and 10dB (Fig. 4).

The notable feature of all these curves is that each has a peak for a value of T which corresponds to a signal design which has neither orthogonal nor identical waveforms.

Note that identical waveform signaling corresponds to a singular correlation matrix $\mathbf{R}$, and thus the total PDD capacity goes to zero as the time T decreases, and the peak capacity of the PDD occurs for nearly orthogonal signaling. These properties, and the gap between the PDD and total OSD capacity curves for intermediate values of T show that the advantage of CWMA is lost by using the PDD.

Figure 5 shows the minimum weighted power required to achieve a total capacity of 4 nats/s as a function of T for FOBP values of .05 and .1. It can be seen here that a large penalty is paid for identical or nearly identical signature waveforms, which occur at the lowest values of T. In Figure 6, we plot the maximum spectral efficiency C/B for fixed B, as a function of SNR, with $\mathbf{A} = \mathbf{I}$. Each point on these curves is maximized over T, with $T \in [.93, 2.8]$ for $\epsilon = .15$, and $T \in [1.29, 3.4]$ for $\epsilon = .05$, and hence represents a different correlation matrix. Further, the OSD and PDD designs are maximized at different values of T for the same FOBP constraint.

VII. CONCLUSIONS

While it is impossible to have signals which are strictly time-limited and strictly bandlimited, we are often faced with the notion of signals which are strictly time-limited and essentially bandlimited. Here, we have used the FOBP bandwidth measure to make the notion of "essentially bandlimited" mathematically precise, and to design signal sets under such a constraint, which is more practical (and general) than a strict bandwidth constraint. By further requiring that the FOBP constraint be location-invariant, we have made this constraint more realistic than in previous papers.

We have obtained signature waveforms and power allocations that optimize total capacity for FOBP-limited channels with CWMA signaling. By such choice of optimally correlated signals and symbol rate, it is possible to have significantly higher capacities (by using the OSD) than with orthogonal or identical signaling. However, these large gains in capacity are not available when the PDD is used as a decoder.

REFERENCES

[1] M. K. Varanasi and T. Guess, "Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, pp. 1405–1409, Nov. 1997.

[2] D. Parsavand and M. K. Varanasi, "RMS bandwidth constrained signature waveforms that maximize the total capacity of PAM-synchronous CDMA channels," *IEEE Trans. Commun.*, vol. 44, pp. 65–75, Jan. 1996.

[3] F. Amoroso, "The bandwidth of digital signals," *IEEE Commun. Mag.*, vol. 18, pp. 13–24, Nov. 1980.

[4] E. A. Lee and D. G. Messerschmitt, *Digital Communication*. Boston, MA: Kluwer Academic Publishers, second ed., 1994.

[5] R. Roberts and C. T. Mullis, *Digital Signal Processing*. Reading, MA: Addison-Wesley, 1987.

[6] M. Hall Jr., *Combinatorial Theory*. Waltham, MA: Blaisdell, 1967.

[7] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press, 1979.

[8] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge University Press, 1985.

[9] D. Slepian and H. O. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty-I," *Bell Syst. Tech. J*, vol. 40, pp. 43–63, 1961.

[10] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*. New York: Dover, 1965.

On the Classification of Formally Self-Dual Codes

J. E. Fields*, P. Gaborit*, W. C. Huffman†, V. Pless*

Abstract

In this paper we discuss the status of the classification of even formally self-dual binary codes. Of most interest are those codes with highest minimum weight. We compare the results on even formally self-dual codes with those of self-dual codes. We give information about both self-dual and even formally self-dual codes for lengths up to 48. A particularly useful form of the generator matrix is presented which was used in the recent classification of the codes of length 14, 20, and 22.

1 Introduction

Self-dual (s.d.) codes are an interesting class of codes. The extended Golay codes over $GF(2)$ and $GF(3)$ are self-dual as are the [8,4,4] binary Hamming code and the $[6, 3, 4]$ hexacode over $GF(4)$. The weight enumerators of these codes are given by the Gleason polynomials which also provide upper bounds on their minimum weights. Codes which meet the highest of these bounds are called *extremal* and vectors of a fixed weight in them contain t-designs where t can be 5, 3, or 1 (depending on the particular case).

In order to find the extremal codes, all self-dual codes of a particular length have been classified for many modest lengths; see [5, 6, 7, 19, 20, 21, 22, 23, 24, 25, 26]. This has been possible because there are formulas for the number of self-dual codes of a fixed length.

Binary self-dual codes come in two types: Type II, the doubly-even codes where all weights are divisible by 4, and Type I, the singly-even codes where some weights are also equivalent to 2 (mod 4). The longest length where all Type II codes have been classified is 32. It is impractical to do this at longer lengths due to the large number of such codes. Attempts have been made (so far unsuccessfully) to classify only the extremal codes at longer lengths even though there are no formulas for the number of such codes. Consequently researchers have looked at certain subfamilies of these extremal codes and successfully classified them; see for example [4, 13, 15, 29, 30, 31, 33, 34]. For more information on these classifications, see [14, 28].

If a binary code has the same weight distribution as its dual code, it is called *formally self-dual (f.s.d.)*. We only consider f.s.d. codes that are *even* (codes with only even weight codewords). These codes include the self-dual codes. When a code has only even weight codewords and the weight enumerator of the code equals the weight enumerator of its dual, this weight enumerator is a combination of the Gleason polynomials [18]. Therefore when the minimum weight is so high that the weight distribution is uniquely

---

*Department of Mathematics, Statistics, and Computer Science, 322 SEO, University of Illinois–Chicago, 851 S. Morgan, Chicago, IL 60607-7045, USA; e-mail: fields@math.uic.edu, gaborit@math.uic.edu, pless@math.uic.edu

†Department of Mathematical and Computer Sciences, Loyola University, Chicago, IL 60626, USA; e-mail: wch@math.luc.edu

n	$d_{s.d.}$	$d_{f.s.d.}$	Type	Number	Dim Hull	Reference
8	4	4	II	1		[23]
10	2^A	4	I, f.s.d.	2, 1	1	[23, 18]
12	4	4	I, f.s.d.	1, 2	2, 4	[23, 1]
14	4	4	I, f.s.d.	1, 9	1, 3, 5	[23, 10]
16	4	4^B	I, II, f.s.d.	1, 2, ≥ 6	2, 4, 6	[23, 12]
18	4	6	I, f.s.d.	2, 1	1	[23, 18, 32]
20	4	6	I, f.s.d.	7, 7	2, 4	[23, 11]
22	6	6	I, f.s.d.	1, > 1000	1, 3, 5, 9, ?	[25, 11, 12]
24	8	8	II	1		[25, 22]
26	6^B	6^B	I, f.s.d.	1, ≥ 30	1, 3, 5, 7, 9, ?	[7, 10, 12]
28	6^A	8	I, f.s.d.	3, 1	2	[7, 18, 16]
30	6^A	8	I, f.s.d.	13, ≥ 6	1, 3, ?	[7, 18, 12]
32	8	8	I, II, f.s.d.	3, 5, ≥ 10	4, 6, 8, 12, 14, ?	[8, 5, 12]
34	6^A	8^B	I, f.s.d.	≥ 200, ≥ 7	1, 9, ?	[8, 12]
36	8	8^B	I, f.s.d.	≥ 4, ≥ 23	2, 4, 6, 8, 10, 12, ?	[34, 12]
38	8	8^B	I, f.s.d.	≥ 7, ≥ 24	1, 3, 5, 17, ?	[31, 12]
40	8	10	I, II, ?	≥ 37, ≥ 1000, ?		[4, 28]
42	8	10	I, f.s.d.	≥ 30, ≥ 6	1, 3, ?	[9, 12]
44	8	10^B	I, f.s.d.	≥ 108, ≥ 12	2, 4, 12, ?	[9, 12]
46	10	10^B	I, f.s.d.	≥ 1, ≥ 41	1, 11, ?	[8, 12]
48	12	12	II, ?	≥ 1, ?		remark 6

Table 1: Highest minimum weights of s.d. and even f.s.d. codes

determined (here, the codes are extremal), even f.s.d. and Type I codes have the same weight distribution. Any Type I code must have the same number of singly-even and doubly-even vectors; however this is not necessarily so for the non-extremal even f.s.d. codes so that they can have weight enumerators which are not possible for Type I codes (see the remark later regarding f.s.d. codes of length 16). Extremal even f.s.d. codes exist which have higher minimum weights than self-dual codes of the same length. Recent work has classified the extremal even f.s.d. codes through length 28.[1]

2 What's Known

Table 1 summarizes the information presently known about s.d. and even f.s.d. codes of length n, with $8 \leq n \leq 48$, and highest possible minimum weight. The columns '$d_{s.d.}$' and '$d_{f.s.d.}$' give upper bounds on the minimum weights of s.d. and even f.s.d. codes, respectively. By [27]

$$d_{s.d.} \leq \begin{cases} 4\lfloor \frac{n}{24} \rfloor + 4 & \text{if } n \not\equiv 22 \pmod{24} \\ 4\lfloor \frac{n}{24} \rfloor + 6 & \text{if } n \equiv 22 \pmod{24} \end{cases} . \tag{1}$$

A bound on $d_{f.s.d.}$ can be obtained as follows. There is a unique weight enumerator for an even f.s.d. $[n, n/2, d]$ code where $d = 2\lfloor \frac{n}{8} \rfloor + 2$. In particular this is the weight enumerator for the even f.s.d. code with largest possible d. Thus in general

---

[1] By [3] there are no $[16, 8, 6]$ and $[26, 13, 8]$ codes of any type and hence no extremal even f.s.d. codes of lengths 16 and 26.

$$d_{f.s.d.} \leq 2\left\lfloor \frac{n}{8} \right\rfloor + 2. \tag{2}$$

However by examining the coefficient A_{d+2} of this weight enumerator, many people have found (see [28, Theorem 29]) that $A_{d+2} < 0$ for

$$n = 8i \ (i \geq 4), \ n = 8i + 2 \ (i \geq 5), \ n = 8i + 4 \ (i \geq 6), \text{ and } n = 8i + 6 \ (i \geq 7). \tag{3}$$

Therefore the bound

$$d_{f.s.d.} \leq 2\left\lfloor \frac{n}{8} \right\rfloor \tag{4}$$

is valid for n in the range given by (3). In particular

$$d_{f.s.d.} \leq 2\left\lfloor \frac{n}{8} \right\rfloor \text{ for } n \geq 56. \tag{5}$$

At times there is no $[n, n/2]$ code of any type whose minimum weight meets one of these bounds. This type of information can be found for example in the tables of A. E. Brouwer (see [3] or his website). Combining Brouwer's tables with (3) and (5) gives

$$d_{f.s.d.} \leq 2\left\lfloor \frac{n}{8} \right\rfloor \text{ for } n \geq 32. \tag{6}$$

In Table 1, we give the appropriate values of $d_{s.d.}$ and $d_{f.s.d.}$ from (1), (2), and (3) except in the following. When Brouwer's tables indicate that no $[n, n/2]$ code of any type exists with this minimum weight, the entries given are the appropriate bound reduced by 2 and are marked as such with a ^{B} in the table. The values for $d_{s.d.}$ with an ^{A} when $n = 10$, 28, 30, and 34 in the table are two below the bound (1) because it can be shown that no s.d. code meeting the bound exists ([7, 8, 23]).

The columns 'Type' and 'Number' indicates the type of code and the number of such codes which meet the appropriate bound. The values given under 'Number' correspond to the codes listed under 'Type'. The entry 'II' refers to doubly-even codes while 'I' refers to self-dual singly-even codes that are not doubly-even; 'f.s.d.' refers to even formally self-dual codes that are not self-dual. Type I and II codes meet the bound $d_{s.d.}$ and even f.s.d. codes meet the bound $d_{f.s.d.}$. The *hull* $\mathcal{H}$ of a code $\mathcal{C}$ is defined to be $\mathcal{H} = \mathcal{C} \cap \mathcal{C}^{\perp}$. The column 'Dim Hull' gives the dimensions of the hulls of the known f.s.d. codes. Entries in 'Dim Hull' occur only if even f.s.d. codes are known to exist. A question mark indicates that the corresponding entry is open.

We make some remarks about certain entries in Table 1.

1. Note that Type II codes can exist only when $n \equiv 0 \pmod 8$. Additionally by [27], when $n \equiv 0 \pmod{24}$, a self-dual code with $d = 4\lfloor \frac{n}{24} \rfloor + 4$ must be Type II.
2. When $n = 8$, the only code meeting the bound is the $[8, 4, 4]$ extended Hamming code. When $n = 24$, the only code meeting the bound is the $[24, 12, 8]$ extended Golay code.
3. When $n = 22$, the authors found 1000 even f.s.d. codes in [11]. These had hull dimensions 1, 3, and 5. In [12], all the $[22, 11, 6]$ even f.s.d. codes double circulant and bordered double circulant codes were found. There were 9 inequivalent codes of these types; they had hull dimensions 1, 5, and 9.

4. In [32], Simonis shows that there is a unique $[18,9,6]$ code. This code turns out to be even f.s.d. (see [18]) and is an extended quadratic residue code. By [16], there is a unique $[28,14,8]$ code; this code is even f.s.d. ([18]).

5. While a $[10,5,4]$ code is not unique, if you add the assumption that the code contains the all-1 codeword, then the weight distribution is unique [17] and this must be the weight distribution of an even f.s.d. code. This code exists and is unique [18].

6. The only $[48,24,12]$ binary code known is the Type II extended quadratic residue code.

7. Note that there exist even f.s.d. codes with higher minimum weight than the s.d. codes when $n = 18$, 20, 28, 30, 34, 42, and 44 (and possibly 40).

This table together with the above remarks leads to the following open problems.

Research Problem 2.1 *Fill in the question marks and decide at which lengths it is possible to classify the even f.s.d. codes.*

Research Problem 2.2 *Is there an even f.s.d.* $[40,20,10]$ *code?*

Research Problem 2.3 *Is there an even f.s.d.* $[48,24,12]$ *code? Is there another Type II* $[48,24,12]$ *code? Most generally, is there another* $[48,24,12]$ *code?*

Research Problem 2.4 *Improve the bound (6) on even f.s.d. codes. (For example, when* $n = 72$*, (6) yields* $d \leq 18$*; however no* $[72,36,18]$ *code of any type exists by [3].)*

3 Methods Used to Find Even F.s.d. Codes

The s.d. codes of highest possible minimum weight, except for lengths 8 and 24, were found by classifying all s.d. codes of the given length. This is possible because there are formulas for the number of all such codes. As no such formulas for f.s.d. codes of a fixed length are known, classifying the even f.s.d. codes which meet the bound '$d_{f.s.d.}$' of Table 1 requires techniques different from those used to classify s.d. codes. The first major development in this problem was the proof of the uniqueness of these codes of lengths 10 and 18 [18]. This relied heavily on the existence of 3-designs held by the codewords of a fixed weight in $C \cup C^{\perp}$, a property that holds for even f.s.d. $[n, n/2, d]$ codes with $d = 2\lfloor \frac{n}{8} \rfloor + 2$ and $n \equiv 2 \pmod 8$. When $n \equiv 6 \pmod 8$, codewords of fixed weight in $C \cup C^{\perp}$ hold only 1-designs, while designs are known to exist in an extremal code C when $n \equiv 0 \pmod 4$ by the Assmus–Mattson Theorem. Only recently has progress been made on classifying these extremal codes for other lengths. This was first done by Bachoc in [1] using intersection numbers (defined below). The current authors have classified the even f.s.d. codes of lengths 14, 20, and 22 [10, 11]. The main tool used is the following form of a generator matrix for these codes. Let $\{X\}$ denote the code with generator matrix X, and k_X the dimension of $\{X\}$.

Theorem 3.1 *Let* C *and* $C^{\perp}$ *be codes of length* $n_1 + n_2$ *with the following generator matrices,* $\mathcal{G}(C)$ *and* $\mathcal{G}(C^{\perp})$*, under the assumption that* A *(respectively* B*) and* F *(respectively*

J) generate the subcodes of $\mathcal{C}$ and $\mathcal{C}^\perp$ of largest dimension with support under the first n_1 (respectively last n_2) coordinates:

$$\mathcal{G}(\mathcal{C}) = \begin{pmatrix} A & 0 \\ 0 & B \\ D & E \end{pmatrix} \quad \textit{and} \quad \mathcal{G}(\mathcal{C}^\perp) = \begin{pmatrix} F & 0 \\ 0 & J \\ L & M \end{pmatrix}.$$

Further assume that $\dim \mathcal{C} = \dim \mathcal{C}^\perp$. *Then*

(a) $k_D = k_E = k_L = k_M$,

(b) $\{A\}^\perp = \{F \cup L\}$, $\{B\}^\perp = \{J \cup M\}$, $\{F\}^\perp = \{A \cup D\}$, $\{J\}^\perp = \{B \cup E\}$, *and*

(c) $n_1 - 2k_A = n_2 - 2k_J$, $n_1 - 2k_F = n_2 - 2k_B$.

Also important is the concept of intersection numbers. Let $\mathcal{C}$ be a binary linear code of length n with dual $\mathcal{C}^\perp$. If $c \in \mathcal{C}$, let supp(c) denote the *support* of c, and let wt(c) denote the *weight* of c. Let t be a positive integer, and let i and w be nonnegative integers. Also let T be a set of t coordinates of the code. The following are *intersection numbers* of $\mathcal{C}$:

$$n_{w,i}(T) = |\{c \in \mathcal{C} \mid \mathrm{wt}(c) = w, \ |\mathrm{supp}(c) \cap T| = i\}|,$$
$$n^*_{w,i}(T) = |\{c \in \mathcal{C}^\perp \mid \mathrm{wt}(c) = w, \ |\mathrm{supp}(c) \cap T| = i\}|.$$

These numbers can often be computed using the generator matrix of Theorem 3.1. (Bachoc used harmonic polynomials to compute them.)

We illustrate how the generator matrix of Theorem 3.1 can be used to compute intersection numbers, in a specific example. In the next section we show how the generator matrix can be used to construct the desired codes. Let $\mathcal{C}$ be an even [20, 10, 6] f.s.d. code. The weight distribution of $\mathcal{C}$ is

$$A_0 = A_{20} = 1, \quad A_6 = A_{14} = 90, \quad A_8 = A_{12} = 255, \quad A_{10} = 332. \tag{7}$$

Let T denote a set of $t = 6$ coordinates and $c(T)$ the vector of length 20 with supp$(c(T)) = T$. In this illustration we only compute $n_{w,i}(T)$. Order the coordinates so that T is the first 6 coordinates and assume $c(T) \in \mathcal{C}$; if $n_1 = 6$, then $A = (111111)$ in Theorem 3.1. There are two possibilities to consider: either $c(T) \in \mathcal{C}^\perp$ or $c(T) \notin \mathcal{C}^\perp$. Here we look only at the first case where $c(T) \in \mathcal{C}^\perp$. The presence of the all-1 vector $\mathbf{1}$ in $\mathcal{H} = \mathcal{C} \cap \mathcal{C}^\perp$ implies that $n_{20-w,t-i}(T) = n_{w,i}(T)$. Hence we only consider $w \in \{6, 8, 10\}$.

Table 2 gives information for the intersection numbers that is easy to obtain using simple coding theory arguments. As $c(T) \in \mathcal{C}^\perp$, $n_{w,i}(T) = 0$ for i odd by orthogonality between $\mathcal{C}$ and $\mathcal{C}^\perp$. Obviously $n_{6,6}(T) = 1$; as $\mathcal{C}$ has minimum weight 6 and maximum weight 14, $n_{6,4}(T) = n_{8,6}(T) = n_{10,0}(T) = n_{10,6}(T) = 0$. By adding $\mathbf{1} + c(T)$, whose support is the 14 coordinates outside of T, to any codeword counted in $n_{6,0}(T)$, we see that $n_{6,0}(T) = n_{8,0}(T)$; similarly $n_{8,2}(T) = n_{10,2}(T)$. By adding $c(T)$ to any codeword counted in $n_{6,2}(T)$, we see that $n_{6,2}(T) = n_{8,4}(T)$; similarly $n_{8,2}(T) = n_{10,4}(T)$.

Thus the only entries among $n_{w,i}(T)$ left to find are x, y, and z. In Theorem 3.1 the code $\{B\}$ is a [14, 5, 6] code which must have weight distribution $a_0 = a_{14} = 1$ and $a_6 = a_8 = 15$. Therefore $x = n_{6,0}(T) = a_6 = 15$. As

$$\sum_{i=0}^{6} n_{6,i}(T) = 90$$

i	0	1	2	3	4	5	6
$n_{6,i}(T)$	x	0	y	0	0	0	1
$n_{8,i}(T)$	x	0	z	0	y	0	0
$n_{10,i}(T)$	0	0	z	0	z	0	0

Table 2: Partial intersection table for $t = 6$ with $c(T) \in \mathcal{C}$ when $n = 20$

by (7), $x + y + 1 = 90$ implying $y = 74$. But by the Assmus–Mattson Theorem, the codewords of weight 6 in $\mathcal{C}$ hold a 1-(20,6,27) design. Therefore

$$\sum_{i=0}^{6} i n_{6,i}(T) = 6 \cdot 27,$$

because the left-hand side counts the total number of times the supports of codewords of weight 6 overlap the set T, and this must be the right-hand side as these codewords support a 1-design. So

$$0 \cdot x + 2 \cdot y + 6 \cdot 1 = 6 \cdot 27,$$

yielding $y = 78$, a contradiction. (We note that using the harmonic polynomials and the methods of Bachoc, we obtain a conflict also.) This proves that the hull of an even $[20, 10, 6]$ f.s.d. code does not contain codewords of weight 6. In particular, this implies that there are no $[20, 10, 6]$ s.d. codes; notice that this is a quick proof of this fact whereas the original proof [23] required the complete classification of all $[20, 10, d]$ s.d. codes.

In [10, 11] we used similar arguments to construct intersection tables for even f.s.d. codes of lengths 14, 20, and 22. In most cases the tables were not unique. As the (unique) even f.s.d. codes of lengths 10 and 18 were constructed in [18] using different techniques without intersection tables, we include in Table 3 information about their intersection numbers. As in the case $n = 20$, it can be shown using intersection numbers that the hull of an even $[10, 5, 4]$ f.s.d. code does not contain codewords of weight 4 and the hull of an even $[18, 9, 6]$ f.s.d. code does not contain codewords of weight 6. Again this shows that there can be no $[10, 5, 4]$ or $[18, 9, 6]$ s.d. codes, consistent with the results in [23]. Table 3 gives the intersection table for a codeword $c(T)$ in an even $[10, 5, 4]$ (respectively $[18, 9, 6]$) f.s.d. code $\mathcal{C}$ such that $c(T) \notin \mathcal{C} \cap \mathcal{C}^{\perp}$ where $t = 4$ (respectively $t = 6$).

4 Code Constructions

We now discuss the construction and classification of even f.s.d. codes using the generator matrix of Theorem 3.1.

In [18], the generator matrix was not used to show that there is a unique even $[10, 5, 4]$ f.s.d. code. We demonstrate here how easy this is to prove using the generator matrix

$[10, 5, 4]$ code with $t = 4$					
i	0	1	2	3	4
$n_{4,i}(T)$	0	8	6	0	1
$n^{*}_{4,i}(T)$	3	0	12	0	0

$[18, 9, 6]$ code with $t = 6$							
i	0	1	2	3	4	5	6
$n_{6,i}(T)$	6	24	39	32	0	0	1
$n_{8,i}(T)$	0	24	42	48	39	0	0
$n^{*}_{6,i}(T)$	12	0	78	0	12	0	0
$n^{*}_{8,i}(T)$	3	0	96	0	54	0	0

Table 3: Intersection tables for $c(T) \in \mathcal{C}$ and $c(T) \notin \mathcal{C}^{\perp}$

in Theorem 3.1 for such a code $\mathcal{C}$. Choose a weight 4 codeword in $\mathcal{C}$ and order the coordinates so that its support is the first 4 positions; hence we can choose A to be the 1×4 matrix (1111) where $n_1 = 4$. As $\mathcal{C}$ contains the all-1 codeword, the length 6 code $\{B\}$ must contain the all-1 vector. Since there is no $[6, 2, 4]$ code with the all-1 vector, $B = (111111)$ and so $\{A \cup D\}$ is 4-dimensional. We can now choose the three rows of D to have weight 1, and the three rows of E must then have weight 3. Up to equivalence, it is easy to see that

$$\mathcal{G}(\mathcal{C}) = \left(\begin{array}{c|c} 1111 & 000000 \\ \hline 0000 & 111111 \\ \hline 1000 & 111000 \\ 0100 & 110100 \\ 0010 & \end{array} \right).$$

The last row of E can now be completed (up to equivalence) by 110010, 101100, or 101010. The first two yield weight 2 codewords and so only the last is possible.

We now discuss the length 20 case. Let $\mathcal{C}$ be an even $[20, 10, 6]$ f.s.d. code. In view of the preceding section we know that no weight 6 vector is in $\mathcal{C}$ and in $\mathcal{C}^\perp$. Let c be a weight 6 codeword in $\mathcal{C}$. Reorder coordinates so that $\text{supp}(c)$ is the first 6 positions. In the notation of Theorem 3.1, $\mathcal{C}$ has a generator matrix $\mathcal{G}(\mathcal{C})$ with $A = (111111)$. As $c \notin \mathcal{C}^\perp$, $k_F = 0$. Therefore D is the matrix given in Figure 1, and $\{B\}$ is a $[14, 4, 6]$ code containing the all-1 vector and hence having weight distribution $a_0 = a_{14} = 1$ and $a_6 = a_8 = 7$. It is easy to prove by hand that there are only two such codes up to equivalence; these have generator matrices $B = B_i$ of Figure 1.

Our most difficult job was to determine the rows of E. Each of these rows can be chosen modulo the code $\{B\} = \{B_i\}$. As $\mathcal{C}$ has minimum weight 6, the rows of E can be chosen from the set of the coset leaders of the code $\{B_i\}$ with odd weight 5 or more. For $\{B_1\}$ there are 200 coset leaders of weight 5 and for $\{B_2\}$ there are 196 coset leaders of weight 5 and 8 of weight 7. Therefore if we apply a brute force method to test all the different possible codes, there are 5 rows of E to determine and therefore about $\frac{200^5}{5!}$ possibilities for each choice of B, if we order them, which is more than 250×10^6 cases. So we had to be more clever and take advantage of the fact that we want to construct $\mathcal{C}$ only up to equivalence. Using group theoretic techniques and Magma [2], we can reduce the number of possibilities; the actual procedure is explained in more detail in [11]. As an illustration, when $B = B_1$, the submatrix of E consisting of the first three rows can be chosen from among only 225 possibilities. In the end the case $B = B_1$ yielded seven inequivalent even f.s.d. codes. Proceeding similarly when $B = B_2$, we found exactly the same seven codes.

We briefly mention the situation in lengths 14 and 22. First let $\mathcal{C}$ be an even $[14, 7, 4]$ f.s.d. code. The classification of [23] shows that there is only one s.d. code of this type.

$$D = \begin{pmatrix} 100000 \\ 010000 \\ 001000 \\ 000100 \\ 000010 \end{pmatrix} \qquad B_1 = \begin{pmatrix} 11001101111000 \\ 00101101100111 \\ 00011100011111 \\ 00000011111111 \end{pmatrix} \qquad B_2 = \begin{pmatrix} 11000000001111 \\ 00100100111100 \\ 00011000110011 \\ 00000011111111 \end{pmatrix}$$

Figure 1: Submatrices in $\mathcal{G}(\mathcal{C})$ for even $[20, 10, 6]$ f.s.d. codes

It is easy to show that if C is not s.d., then C contains a codeword of weight 4 that is not in $C^{\perp}$. Using the support of such a codeword as the first 4 coordinate positions and arguing as in the length 20 case, we find that there are two inequivalent choices for B and that E is a 3×10 matrix. A brute force method works quickly to find E. For each choice of B we find the same nine inequivalent non-self-dual even f.s.d. codes.

Length 22 proved to be very surprising. We assumed that there was a codeword in C of weight 6 that is not in $C^{\perp}$. Using the support of this codeword as the first 6 coordinate positions, thus determining the matrices A and D, we find that there are eleven inequivalent choices for B. For each choice of B, the rows of the 5×16 matrix E must be chosen from approximately 500 coset leaders. As we began to generate codes, the number of inequivalent even $[22, 11, 6]$ f.s.d. codes grew rapidly. We stopped generating them after we reached 1000 codes; we had only tested a small fraction of the total possibilities and so suspect that the actual number is much higher.

Some interesting observations have arisen in our study of f.s.d. codes. We describe some of these.

- All the extremal even f.s.d. codes of lengths through 12 have (pure) double circulant generator matrices. At length 14, only 2 of the 10 even f.s.d. codes have such generator matrices.

- All of the extremal even f.s.d. codes of length through 20 are isodual (equivalent to their dual).[2] The first non-isodual codes were found at length 22. In fact of our 1000 codes, 963 were not isodual!
- Up through length 20, no extremal even f.s.d. codes have trivial automorphism group. (At length 20, the automorphism groups were relatively small with orders 8, 9, 10, and 20.) However at length 22, codes with trivial groups seem to be abundant. Of the 1000 codes found, 972 had trivial automorphism group. Only two had groups of order larger than 4.

- When examining Table 1, we notice a pattern in the parity of the dimensions of the hulls. This pattern in fact holds in general; in [10] it is proved that if C is an even f.s.d. code of length n, then the dimension of the hull is odd if $n \equiv 2 \pmod 4$ and is even if $n \equiv 0 \pmod 4$.

References

[1] C. Bachoc, "On harmonic weight enumerators of binary codes," preprint.

[2] W. Bosma and J. Cannon, *Handbook of Magma Functions*, Sydney, 1995.

[3] A. E. Brouwer, "Bounds on the size of linear codes," *Handbook of Coding Theory*, edited by V. Pless and W. C. Huffman, North Holland, Amsterdam, 1998.

[4] S. Buyuklieva and V. Y. Yorgov, "Singly-even self-dual codes of length 40," *Designs, Codes and Crypt.* **9** (1996), 131–141.

[5] J. H. Conway and V. Pless, "On the enumeration of self-dual codes," *J. Comb. Theory* **A 28** (1980), 26–53.

---

[2] Note that the even $[16, 8, 4]$ f.s.d. codes are not extremal.

[6] J. H. Conway, V. S. Pless, and N. J. A. Sloane, "Self-dual codes over GF(3) and GF(4) of length not exceeding 16," *IEEE Trans. Inform. Theory* **IT-25** (1979), 312–322.

[7] J. H. Conway, V. S. Pless, and N. J. A. Sloane, "The binary self-dual codes of length up to 32: a revised enumeration," *J. Comb. Theory* **A 60** (1992), 183–195.

[8] J. H. Conway and N. J. A. Sloane, "A new upper bound on the minimal distance of self-dual codes," *IEEE Trans. Information Theory* **36** (1990), 1319-1333.

[9] S. T. Dougherty, T. A. Gulliver, and M. Harada, "Extremal binary self-dual codes," *IEEE Trans. Information Theory* **43** (1997), 2036–2047.

[10] J. E. Fields, P. Gaborit, W. C. Huffman, and V. Pless, "On the classification of extremal even formally self-dual codes," preprint.

[11] J. E. Fields, P. Gaborit, W. C. Huffman, and V. Pless, "On the classification of extremal even formally self-dual codes of lengths 20 and 22," preprint.

[12] T. A. Gulliver and M. Harada, "Classification of extremal double circulant formally self-dual even codes," *Designs, Codes and Crypt.* **11** (1997), 25–35.

[13] W. C. Huffman, "Automorphisms of codes with applications to extremal doubly even codes of length 48," *IEEE Trans. Inform. Theory* **IT-28** (1982), 511–521.

[14] W. C. Huffman, "On the classification of self-dual codes," *Proc. 34th Allerton Conf. Commun. Control and Computing (October 2-4, 1996)*, 302–311.

[15] W. C. Huffman and V. D. Tonchev, "The existence of extremal [50, 25, 10] codes and quasi-symmetric 2-(49,9,6) designs," *Designs, Codes and Crypt.* **6** (1995), 97–106.

[16] D. B. Jaffe, "Optimal binary linear codes of lengths ≤ 30," *Discrete Math*, to appear.

[17] D. B. Jaffe, private communication.

[18] G. T. Kennedy and V. Pless, "On designs and formally self-dual codes," *Designs, Codes and Cryptography* **4** (1994), 43–55.

[19] J. S. Leon, V. Pless, and N. J. A. Sloane, "On ternary self-dual codes of length 24," *IEEE Trans. Inform. Theory* **IT-27** (1981), 176–180.

[20] F. J. MacWilliams, A. M. Odlyzko, N. J. A. Sloane, and H. N. Ward, "Self-dual codes over GF(4)," *J. Combin. Theory A* **25** (1978), 288–318.

[21] C. L. Mallows, V. Pless, and N. J. A. Sloane, "Self-dual codes over GF(3)," *SIAM J. Appl. Math.* **31** (1976), 649–666.

[22] V. Pless, "On the uniqueness of the Golay codes," *J. Comb. Theory* **5** (1968), 215–228.

[23] V. Pless, "A classification of self-orthogonal codes over GF(2)," *Discrete Math.* **3** (1972), 209–246.

[24] V. Pless, "The children of the (32,16) doubly even codes," *IEEE Trans. Inform. Theory* **IT-28** (1982), 738–746.

[25] V. Pless and N. J. A. Sloane, "On the classification and enumeration of self-dual codes," *J. Comb. Theory* **18** (1975), 313–335.

[26] V. Pless, N. J. A. Sloane, and H. N. Ward, "Ternary codes of minimum weight 6 and the classification of the self-dual codes of length 20," *IEEE Trans. Inform. Theory* **IT-26** (1980), 305–316.

[27] E. M. Rains, "Shadow bounds for self-dual codes," *IEEE Trans. Inform. Theory* **44** (1998), 134–139.

[28] E. M. Rains and N. J. A. Sloane, "Self-dual codes," *Handbook of Coding Theory*, edited by V. Pless and W. C. Huffman, North Holland, Amsterdam, 1998.

[29] R. P. Ruseva, "Uniqueness of the [36, 18, 8] double circulant code," in *Proc. of the International Workshop on Optimal Codes and Related Topics*, May 26–June 1, 1995, Sozopol, Bulgaria, 126–129.

[30] R. P. Ruseva, "New extremal self-dual codes of length 36," in *Proc. of the* 25th *Spring Conf. of the UBM*, 1996, 150–153.

[31] R. P. Ruseva, "On the extremal self-dual binary codes of length 38 with an automorphism of order 7," preprint.

[32] J. Simonis, "The [18,9,6] code is unique," *Discrete Math.* **106/107** (1992), 439–448.

[33] V. Y. Yorgov, "Binary self-dual codes with automorphisms of odd order," *Prob. Inform. Trans.* **19** (1983), 11–24.

[34] V. Yorgov and N. Yankov, "On the extremal binary codes of lengths 36 and 38 with an automorphism of order 5," *Proc. of the* 5th *International Workshop on Algebraic and Combinatorial Coding Theory* June 1–7, 1996, Sozopol, Bulgaria, 307–312.

On the Dual of MDS Convolutional Codes*

Joachim Rosenthal, Roxana Smarandache

Department of Mathematics
University of Notre Dame
Notre Dame, IN 46556.
e-mail: Rosenthal.1@nd.edu, Smarandache.1@nd.edu
URL: http://www.nd.edu/~eecoding/

September 21, 1998

Abstract

A maximum distance separable (MDS) convolutional code is a linear code whose distance is maximal among all linear convolutional codes of rate k/n and fixed degree δ. We have conjectured already in previous work that MDS convolutional codes do exist if one allows sufficiently many field elements. In this paper we will give the proof of this fact in the case $\delta = 0$ without making use of the very well known Reed-Solomon construction. This will illustrate how things work in the general case. Also we will question the nature of the dual code of a MDS-convolutional code. We will show that in general the dual of a MDS-convolutional code is not MDS.

Keywords: MDS codes, convolutional codes.

1 Introduction

In the literature there were already several papers [3, 8] which considered the concept of a maximum distance separable convolutional code. In each of these approaches it was necessary to restrict the total class of rate k/n convolutional codes to a suitable subclass. This is simply due to the fact that there is in general no upper bound for the free distance of a rate k/n convolutional code.

Following the module theoretical approach of [9] we define a rate k/n convolutional code $\mathcal{C}$ over a field $\mathbb{F}$ as a rank k submodule of $\mathbb{F}^n[z]$. We argue that the single most important parameter for a k/n convolutional code is *the degree* or *the complexity* (see e.g. [7]). The set of all convolutional codes of rate k/n and degree (complexity) at most δ forms a finite set and consequently the free distances of these codes have to be bounded from above. The generalized Singleton bound which we derived in [10, 11] has

---

*Both authors were supported in part by NSF grant DMS-96-10389.

the property that every convolutional code of rate k/n and degree δ has a free distance of less than this bound and the existence of codes which achieve this distance made appropriate the definition of MDS-convolutional codes.

The degree of the code is easily computed from the code $\mathcal{C}$ directly. For this let $G(z)$ be a $k \times n$ polynomial matrix whose rows generate the code $\mathcal{C}$. In terms of $G(z)$ the degree is equal to the maximal degree of the $k \times k$ full size minors of $G(z)$. (See [6, 9] for details). Note that due to our definition the degree is independent of the particular choice of generator matrix and in this way we can speak of the degree of a code. This point of view differs slightly from the standard one as presented in [1, 6, 7].

Let ν_i be the degree of the ith row of $G(z)$. I.e. $\nu_i = \max_j \deg g_{ij}(z)$. We denote by G_∞ the high order coefficient matrix of $G(z)$. In general G_∞ has not full rank k. Every code $\mathcal{C}$ of rank k has however a $k \times n$ generator matrix $G(z)$ whose row degrees $\nu_1, \dots, \nu_k$ are non-increasing and whose high order coefficient matrix G_∞ has rank k. The degree δ is in this case equal to $\delta = \sum_i \nu_i$ and we say that $G(z)$ is in *row proper form*. The ordered indices $\nu_1 \geq \cdots \geq \nu_k$ are invariants of the convolutional code and we call these indices the *constraint length* of the convolutional code $\mathcal{C}$.

For any n-component vector $v \in \mathbb{F}^n$, we define its weight and denote it by $\mathrm{wt}(v)$, the number of all its nonzero components. The weight of a polynomial with coefficients in $\mathbb{F}^n$ is then the sum of the weights of all its coefficients. Finally we define the free distance of the convolutional code $\mathcal{C} \subset \mathbb{F}^n[z]$ through:

$$d_{free} = \min\{\mathrm{wt}(v(z)) \mid v(z) \in \mathcal{C}, v(z) \neq 0\}.$$

The paper is structured as follows: In Section 2 we give the definition of the MDS-convolutional codes, with the particular construction of Justesen [2] in the rate $1/n$-case. A theoretic proof of the case $\delta = 0$ without making use of the Vandermonde parity check matrix construction is detailed in the third section of the paper. This approach constitutes a base for the general case of rate k/n as it is presented in [10]. In the last section we present some example of MDS-convolutional codes which have a non-MDS dual code.

2 Main Results

Let $\mathcal{C}$ be a convolutional code of rate k/n and degree δ and let G be a polynomial encoder with ordered row degrees $\nu_1 \geq \cdots \geq \nu_k$. We have the following upper bound on the free distance of the code:

Lemma 2.1 *Let r be the number of indices ν_i among the ordered indices $\nu_1 \geq \cdots \geq \nu_k$ having the value $\nu_i = \nu_k$. Then the free distance must satisfy*

$$d_{free} \leq n(\nu_k + 1) - r + 1. \tag{2.1}$$

The set of rate k/n convolutional codes of degree δ is partitioned into sets of codes with different constraint length $\nu_1 \geq \cdots \geq \nu_k$. The upper bound (2.1) is largest if ν_k

is as large as possible and r as small as possible. The largest possible value for ν_k is $\nu_k = \lfloor \delta/k \rfloor$. In this situation the constraint length must have the values

$$\nu_1 = \lfloor \delta/k \rfloor + 1, \cdots, \nu_{k-l} = \lfloor \delta/k \rfloor + 1, \nu_{k-l+1} = \lfloor \delta/k \rfloor, \cdots, \nu_k = \lfloor \delta/k \rfloor.$$

The above set of indices are sometimes referred to as the generic set of row indices. Specializing the above result to the 'generic set of row degrees' we get the following upper bound in terms of the degree δ:

Theorem 2.2 *For every base field $\mathbb{F}$ and every rate k/n convolutional code C of degree δ, the free distance is bounded by:*

$$d_{free} \leq (n-k)\left(\lfloor \delta/k \rfloor + 1\right) + \delta + 1. \tag{2.2}$$

The main Theorem 2.4 will state that there exist always rate k/n convolutional codes of degree δ whose free distance is equal to the right hand side of (2.2). Based on this we define:

Definition 2.3 A rate k/n code of degree δ whose free distance achieves the upper bound (2.2) will be called a MDS convolutional code.

The main result of this paper now states:

Theorem 2.4 *For any rate k/n and any degree δ there exist MDS convolutional codes for sufficiently large field sizes.*

The proof of the main theorem is quite involved and it is given in [10]. The theorem generalizes several special known situations. If $\delta = 0$ the bound (2.2) coincides with the Singleton bound (see e.g. [5]). In this situation we therefore have:

Lemma 2.5 *If G is a $k \times n$ generator of a MDS block code then G also generates a MDS convolutional codes of rate k/n, degree $\delta = 0$ and free distance $n - k + 1$. In particular if $|\mathbb{F}| \geq n$ MDS convolutional codes of rate k/n and degree 0 do exist.*

If $k = 1$ convolutional codes which are MDS were studied by Justesen [2]. In this situation the upper bound (2.2) reduces to

$$d_{free} \leq n(\delta + 1) \tag{2.3}$$

and Justesen did show:

Theorem 2.6 (Justesen [2]) *For any rate $1/n$ and $n \leq q - 1$, the generator polynomials*

$$g_1 = \prod_{k=1}^{\delta}(z - \alpha^k), \quad g_j(z) = g_1(z\alpha^{-s_j}), 2 \leq j \leq n$$

with $|\mathbb{F}| = q$ and $s_j = \lceil (j-1)(q-1)/n \rceil$, define a q-ary convolutional code with free distance

$$d_{free} = n(\delta + 1)$$

when

$$\delta/(q-1) \leq \frac{2kn - k(k+1) + 2f(m)}{n(2k+m)}$$

for all m,

$$f(m) = \begin{cases} (5/8)m^2(1/4)m, & for \quad m \quad even \\ (5/8)m^2(5/8), & for \quad m \quad odd \end{cases}$$

and k is the largest integer such that

$$k/n < 1 - \delta/(q-1).$$

3 The proof idea of Theorem 2.4 in the situation where the degree $\delta = 0$

We will show how the proof goes for the block code situation in order to illustrate some major ideas of the general case.

Let $\mathbb{F}$ be a fixed finite field, with q elements, having characteristic p. Let $\mathbb{K}$ denote the algebraic closure of $\mathbb{F}$. As an algebraically closed field $\mathbb{K}$ is infinite.

We know that all block codes of rate k/n can be described by a parity check matrix of sizes $(n-k) \times n$. We will prove the existence of a $(n-k) \times n$-matrix with all the full size minors invertible, without making use of the Vandermonde-parity check matrix construction. This will result into a MDS block code. Thus let $\mathbf{j} = (1 \leq j_1 < j_2 < \ldots < j_{n-k} \leq n)$ be a fixed index set. We look at the sets $S_{\mathbf{j}}$ of all $(n-k) \times n$ matrices having the property that the minor formed by the $j_1, j_2, \ldots, j_{n-k}$ columns is invertible. We will identify a parity check matrix M with a point in the vector space $\mathbb{K}^{n(n-k)}$. Consider the polynomial ring $\mathbb{K}[X_{ij}]$, $1 \leq i \leq n-k$, $1 \leq j \leq n$. Let $P_{\mathbf{j}} \in \mathbb{K}[X_{ij}]$ be the polynomial $P_{\mathbf{j}} = \det(X_{i,j_s})_{1 \leq i,\ s \leq n-k}$. In other words $P_{\mathbf{j}} = 0$ if and only if the determinant of the minor formed by the columns $j_1, j_2, \ldots, j_{n-k}$ is zero. Let $S_{\mathbf{j}}$ be the Zariski open set defined by

$$S_{\mathbf{j}} = \{M \in \mathbb{K}^{(n-k)n} \mid P_j(M) \neq 0\}.$$

All of these sets $S_{\mathbf{j}}$ form a finite number of open sets in $\mathbb{K}^{(n-k)n}$, all of them nonempty. Indeed, for each $\mathbf{j} = (1 \leq j_1 < j_2 < \ldots < j_{n-k} \leq n)$ we choose a matrix having the identity matrix I_{n-k} on the $j_1, j_2, \ldots, j_{n-k}$ columns and this shows that $S_{\mathbf{j}} \neq \emptyset$. Therefore, as a result of algebraic geometry asserts, the intersection of all the sets $S_{\mathbf{j}}$ is nonempty inside $\mathbb{K}^{(n-k)n}$. It implies the existence of a matrix M having the properties of all the sets $S_{\mathbf{j}}$, hence all its full size minors are invertible.

So far we obtained a matrix $M \in \mathbb{K}^{(n-k)n}$ having the components in $\mathbb{K}$, the algebraic closure of $\mathbb{F}$, and laying in the intersection of all the sets $S_{\mathbf{j}}$. Since the extension $\mathbb{F} \subset \mathbb{K}$ is algebraic, it implies that every component of M is algebraic over $\mathbb{F}$, therefore also finite. If we denote with m_{ij} the components of M we have that all $m_{ij} \in \mathbb{F}[m_{ij}, 1 \leq i \leq n-k, 1 \leq j \leq n]$, which is a finite extension over $\mathbb{F}$, therefore is finite of degree say m. Therefore the code having M as a parity check matrix will be a MDS code over a finite field $\mathbb{F}_{q^m}$, with m possibly rather large.

Of course in the block code case we know already that taking M to be Vandermonde

$$M = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & x & \cdots & x^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x^{(n-k-1)} & \cdots & x^{(n-1)(n-k-1)} \end{pmatrix};$$

with x a primitive element of the field $\mathbb{F} = \mathbb{F}_q$, we obtain a MDS block code. But in the situation of convolutional codes with the degree $\delta > 0$ we do not have such an explicit construction. Therefore the above proof becomes valuable.

In the general case $\delta > 0$, we exhibit a parameterization on the set of all rate k/n convolutional codes of degree δ using a large $\mathbb{F}$-vector space. Then we show that the set of codes which are not MDS forms an algebraic subset. For the parameterization we will use Theorem 3.1 called the generalized first order representation as studied in the systems literature [4].

Theorem 3.1 *Assume $\mathcal{C} \subset \mathbb{F}^n[z]$ is a rate k/n convolutional code of degree δ. Then there exist matrices $K, L \in \mathbb{F}^{(n-k+\delta)\times\delta}$ and $M \in \mathbb{F}^{(n-k+\delta)\times n}$ such that:*

$$\mathcal{C} = \mathcal{C}(K, L, M) = \{v(z) \in \mathbb{F}^n[z] \mid \exists x(z) \in \mathbb{F}^\delta[z] : (zK + L)x(z)^t + Mv(z)^t = 0\}. \quad (3.1)$$

Moreover, the following conditions are satisfied:

1. K has full column rank;
2. $(K|M)$ has full row rank;
3. rank $(z_0K + L \mid M) = \delta + n - k, \forall z_0 \in \mathbb{K}$.

Let V be the vector space of all matrices K, L, M having sizes $(n-k+\delta)\times\delta$, $(n-k+\delta)\times\delta$ and $(n-k+\delta)\times n$ respectively. Inside V it is possible to describe the set of MDS convolutional codes again by a finite intersection of nonempty Zariski open sets. Having done this, the proof of the main theorem in the general case follows the line of the block code situation. Details can be found in [10].

4 The dual of an MDS-convolutional code

After generalizing the notion of MDS block code it naturally arises the question on the nature of the dual code. We know that a dual of a MDS-block code is MDS, so we want to find out if this generalizes to the case of general convolutional codes with degree $\delta > 0$. In the sequel we cover two special cases where this turns out to be true and then we give two examples of MDS-convolutional codes whose dual is not MDS.

In order to introduce the notion of a dual convolutional code in our module theoretical setting, consider the following non-degenerate inner product:

$$\begin{array}{rcl} <,>: \mathbb{F}^n[z] \times \mathbb{F}^n[z] & \longrightarrow & \mathbb{F}[z] \\ (v(z), w(z)) & \longmapsto & v(z)w(z)^t. \end{array} \quad (4.1)$$

Using this inner product we define the dual of a code $\mathcal{C}$ as

$$\mathcal{C}^{\perp} := \{w(z) \mid < v(z), w(z) >= 0, \forall v(z) \in \mathcal{C}\}.$$

One always has that

$$\mathcal{C}^{\perp\perp} \supseteq \mathcal{C}.$$

If the code $\mathcal{C}$ has a minimal basis encoder (i.e. it is non-catastrophic) then $\mathcal{C}^{\perp\perp} = \mathcal{C}$.

The following two lemmas cover some cases where the dual of a MDS convolutional code is MDS.

Lemma 4.1 *If $\mathcal{C}$ is a convolutional code of degree $\delta = 0$ then $\mathcal{C}$ is MDS if and only if $\mathcal{C}^{\perp}$ is MDS.*

Lemma 4.2 *Assume $k = 1, n = 2$. A non-catastrophic code $\mathcal{C}$ of rate $1/2$ is MDS if and only if $\mathcal{C}^{\perp}$ is MDS.*

Following the approach of Justesen, we will give here a concrete example of a 1/3 MDS-convolutional code that has its dual code non-MDS. Then we will present another example of the same type, in the case $\delta = 1 < k$.

Example 4.3 Let α be a primitive element in the field $\mathbb{F}_8$, satisfying $\alpha^3 + \alpha + 1 = 0$. Then the code of degree 2 given by the generator matrix

$$G = \begin{pmatrix} g_1 & g_2 & g_3 \end{pmatrix} = \begin{pmatrix} (z-1)(z-\alpha) & (z-\alpha)(z-\alpha^2) & (z-\alpha^2)(z-\alpha^3) \end{pmatrix}$$

is a MDS-convolutional code. The dual of this code in not MDS.

Proof: For the proof we remind the weight retaining property of Justesen [3]:

Weight-Retaining Property: For any polynomial $T(z)$, any nonzero α in $\mathbb{F}$, any non-negative integers n and N,

$$\mathrm{wt}[T(z)(z^n - \alpha)^N] \geq \mathrm{wt}[(z-\alpha)^N]\ \mathrm{wt}[T(z) \bmod (z^n - \alpha)].$$

We denote the message by $i(z)$. Then the weight of any codeword $v(z) = i(z)G(z)$ is

$$\mathrm{wt}[v(z)] = \mathrm{wt}[g_1(z)i(z)] + \mathrm{wt}[g_2(z)] + \mathrm{wt}[g_3(z)].$$

First we factor out any multiple of $(z^7 - 1)$ that divides $i(z)$. We have:

$$i(z) = (z^7 - 1)^r i'(z)$$

where $z^7 - 1$ does not divide $i'(z)$ anymore. We also denote by h_1, h_2, h_3 the dual polynomials of g_1, g_2, g_3. We have now three cases:

1. If none of the dual generators divides $i'(z)$. Applying the Weight-Retaining Property we obtain:

$$\begin{aligned} \mathrm{wt}[v(z)] \geq\ & \mathrm{wt}[(z-1)^r]\mathrm{wt}[i'(z)g_1(z) \bmod (z^7 - 1)] \\ & + \mathrm{wt}[(z-1)^r]\mathrm{wt}[i'(z)g_2(z) \bmod (z^7 - 1)] \\ & + \mathrm{wt}[(z-1)^r]\mathrm{wt}[i'(z)g_3(z) \bmod (z^7 - 1)] \geq 3(2+1) = 9. \end{aligned}$$

2. If only one of the dual generators, say h_1, divides $i'(z)$: $i'(z) = h_1(z)p(z)$, we have:

$$\begin{aligned}\mathrm{wt}[v(z)] &\geq \mathrm{wt}[(z-1)^{r+1}\mathrm{wt}[(p(z)\mathrm{mod}\,(z^7-1)]\\ &\quad + \mathrm{wt}[(z-1)^r]\mathrm{wt}[p(z)h_1(z)g_2(z)\mathrm{mod}\,(z^7-1)]\\ &\quad + \mathrm{wt}[(z-1)^r]\mathrm{wt}[p(z)h_1(z)g_3(z)\mathrm{mod}\,(z^7-1)] \geq 2+7+6 \geq 9.\end{aligned}$$

3. If two of the duals , say h_1 and h_2, divide $i'(z)$, we have:

$$\begin{aligned}\mathrm{wt}[v(z)] &= \mathrm{wt}[(z-1)^{r+1}]\mathrm{wt}[p(z)\ gcd\ (g_1(z),\ h_2(z))\mathrm{mod}\,(z^7-1)]\\ &\quad + \mathrm{wt}[(z-1)^{r+1}]\mathrm{wt}[p(z)\ gcd\ (g_2(z),\ h_1(z))\mathrm{mod}\,(z^7-1)]\\ &\quad + \mathrm{wt}[(z-1)^r]\mathrm{wt}[p(z)g_3(z)\ lcm(h_1,\ h_2)\mathrm{mod}\,(z^7-1)] \geq 2+2+7 \geq 9.\end{aligned}$$

We have then that the weight of any codeword is greater or equal than $9 = 3(\delta+1)$, therefore the code generated by G is a $1/3$-MDS convolutional code.

Looking now at the dual code, we need to find a 2×3 polynomial matrix having the full size minors equal to the full size minors of the matrix G. By inspection we obtain that a generator matrix for the dual code is

$$G' = \begin{pmatrix} (z-\alpha^2) & (z-1) & 0 \\ 0 & (z-\alpha^3) & (z-\alpha) \end{pmatrix}.$$

The code just obtained has a codeword of weight 4 strictly less than the upper bound of (2.2). Therefore the dual code is not MDS. □

Example 4.4 Consider $k = n-1$, $\delta = 1$ and

$$G(z) = \begin{bmatrix} g_0 + g_1 z \\ G_2 \end{bmatrix}.$$

Here g_0 and g_1 are constant row vectors and G_2 is a matrix of size $(n-2)\times n$. If $G(z)$ is non-catastrophic one shows that $G(z)$ is MDS if and only if

1. G_2 is MDS when considered as the generator matrix of a block code.
2. $\begin{bmatrix} g_0 \\ G_2 \end{bmatrix}$ is MDS or $\begin{bmatrix} g_1 \\ G_2 \end{bmatrix}$ is MDS.

The dual code however is MDS if and only if $\begin{bmatrix} g_0 \\ G_2 \end{bmatrix}$ and $\begin{bmatrix} g_1 \\ G_2 \end{bmatrix}$ are MDS, which are different algebraic conditions.

The above two example show that in general the dual code of a MDS-convolutional code is not an MDS convolutional code anymore in contrast to the situation of block codes.

References

[1] G. D. Forney. Convolutional codes I: Algebraic structure. *IEEE Trans. Inform. Theory*, IT-16(5):720–738, 1970.

[2] J. Justesen. An algebraic construction of rate $1/\nu$ convolutional codes. *IEEE Trans. Inform. Theory*, IT-21(1):577–580, 1975.

[3] J. Justesen and L.R. Hughes. On maximum-distance-separable convolutional codes. *IEEE Trans. Information Theory*, IT-20:288, 1974.

[4] M. Kuijper. *First-Order Representations of Linear Systems.* Birkhäuser, Boston, 1994.

[5] F. J. MacWilliams and N. J.A. Sloane. *The Theory of Error-Correcting Codes.* North Holland, Amsterdam, 1977.

[6] R.J. McEliece. The algebraic theory of convolutional codes. In R. Brualdi, W.C. Huffman, and V. Pless, editors, *Handbook of Coding Theory.* Elsevier Science Publishers, Amsterdam, The Netherlands, 1998. To appear.

[7] Ph. Piret. *Convolutional Codes, an Algebraic Approach.* MIT Press, Cambridge, MA, 1988.

[8] Ph. Piret and T. Krol. MDS convolutional codes. *IEEE Trans. Inform. Theory*, 29(2):224–232, 1983.

[9] J. Rosenthal, J. M. Schumacher, and E.V. York. On behaviors and convolutional codes. *IEEE Trans. Inform. Theory*, 42(6):1881–1891, 1996.

[10] J. Rosenthal and R. Smarandache. Maximum distance separable convolutional codes. In preparation, 1998.

[11] R. Smarandache and J. Rosenthal. Convolutional code constructions resulting in maximal or near maximal free distance. In *Proceedings of the 1998 IEEE International Symposium on Information Theory*, page 308, Boston, MA, 1998.

Algebraic Geometry Codes
are
generalized Reed-Solomon Codes

Douglas A. Leonard
Department of Discrete and Statistical Sciences,
Math Annex, Auburn University, Auburn, AL 36849-5307
leonada@mail.auburn.edu

Abstract— AG codes should be viewed as ideals in multivariate polynomial rings, directly generalizing Reed-Solomon codes, which are, after all, ideals in univariate polynomial rings. If RS codes are properly viewed, then AG codes and their decoding schemes can be easily *rediscovered* as generalizations of RS codes.

I. INTRODUCTION

There have been many recent articles on how to view AG codes [19], [9], [10], [5]. This is yet another, with a slightly different perspective– namely that polynomial orderings in multivariate polynomial rings are more important than algebraic geometry. Thus multivarite polynomial ring ideal theory will replace most of the algebraic geometry. ([2] is a good introduction to ideals, monomial orderings, and varieties.) There will also be parallels drawn to row-reduction and shifting, the basic operations of the Berlekamp-Massey and Sakata algorithms.

II. Reed-Solomon Codes

Let $\mathbb{F}(q)$ denote the algebraic closure of the finite field $GF(q)$. Let α be a primitive element of $GF(q)$ (and hence a primitive $(q-1)$-st root of 1). Consider the Reed-Solomon codes (of length $n = q-1$) with generator polynomials:

$$g(x) := \prod_{i=0}^{r-1}(x - \alpha^i).$$

This has an $r \times n$ parity-check matirx H with entries

$$H_{i,j} := \alpha^{ij}$$

and syndromes

$$s_i(\underline{e}) := (\underline{e}H^T)_i = \sum_j e_j \alpha^{ij}.$$

The standardly used syndrome matrix is

$$S := H\Delta(\underline{e})H^T,\ \Delta(\underline{e}) := diag(\underline{e}),$$

with entries

$$S(\underline{e})_{i,j} = \sum_k H_{i,k}e_k H_{j,k} = \sum_k e_k \alpha^{ik}\alpha^{jk} = s_{i+j}(\underline{e}).$$

The error-locator $\sigma(x) := \prod\{x - a^j : e_j \neq 0\}$ gives rise to functions

$$\sigma_k(x) := \sigma(x)/(x - a_k) =: \sum_l \sigma_{k,l}x^l,$$

zero at all the error locations except a_k, which can be used in *Forney's formula*

$$e_k\left(\sum_l \sigma_{k,l}(a^k)^l\right) = \left(\sum_l \sigma_{k,l}s_l(\underline{e})\right)$$

to produce the error magnitudes.

The first thing we need to do is view this parity-check matrix H as having rows indexed by the functions x^i and columns indexed by the elements α^j. (This is, after all, the view taken in *functional encoding* when H is used as a generator matrix for the dual RS code.) Then think of the set $\Delta_r := \{x^0, x^1, \ldots, x^{r-1}\}$ as the delta set of the ideal $I_r := \langle x^r \rangle$ of the (univariate) polynomial ring $\mathbb{F}(q)[x]$; that is, the set of monomials that are *not* leading monomials of any elements of the ideal. Note that the functions x^r are naturally ordered, so that the delta sets Δ_r and the ideals I_r are naturally nested sequences.

Next, *at least theoretically*, consider the affine point α^j as a projective point $(1 : \alpha^j)$ (on the projective line $\mathbb{P}^1(q)$). And consider the (affine) function x^i as a rational (homogeneous) function $(x_1/x_0)^i$, to realize that $(x_1/x_0)^i$ is a function with no poles on $\mathbb{P}^1(q)$ except for a pole of order i at $P_\infty := (0 : 1)$, and having value a^i at any other point $(1 : a)$ of $\mathbb{P}^1(q)$. Note also that these functions, delta sets, and ideals don't really depend on a particular field (or on using specific points from a specific field), so that they could be evaluated at *any* n points of the projective line.

III. IDEAL CODES

Now generalize $\mathbb{F}(q)[x]$ to the multivariate polynomial ring $\mathbb{F}(q)[x_1, x_2, \ldots, x_N]$. The powers of x were naturally ordered. So order the monomials of this ring so that each monomial is preceded by at most finitely many others. Most symbolic manipulation packages today have the options of weighted (grevlex) orderings, slightly more general orderings defined by a non-singular non-negative integer matrix, or even more general user-defined orderings. It suffices for now to consider the weighted grevlex orderings (possibly, by giving their matrix definition). The variables may come with some natural weights (for instance, their pole orders). Whether they do or not, assign a positive weight w_k to the variable x_k. The matrix that describes the corresponding weighted grevlex monomial ordering is M with $M_{0,k} := w_k$, $M_{l,k} := 1$ iff $k \geq l$, $l \geq 1$, and $M_{l,k} := 0$ otherwise. Then $\underline{x}^{\underline{a}} \succ \underline{x}^{\underline{b}}$ iff $\underline{a}M \succ_{lex} \underline{b}$. (This ordering prefers monomials of larger weight, and then breaks ties by slighting those with more x_1's, etc.)

Now define a sequence of sets as Δ_r being the set of the first r monomials in the given ordering. Let I_r be a sequence of ideals having the corresponding Δ_r as their delta sets; that is the monomials that are not leading monomials of any elements of I_r. Each Δ_r can be used as an index set for a parity-check matrix of a code C_r, (given some choice of n projective points at which all r monomials can be evaluated). The choice of such points is generally all of the projective points defined over a particular finite field $GF(q)$ that are zeros of some (possibly empty) ideal I of functions. If there are actually functions in the ideal, then the variables are not independent, so the delta-set Δ_r should be altered to contain the first r *linearly independent* monomials. (It should be noted that the approach of [5] uses different choices of delta sets to produce *improved* AG codes.)

Example 1 *Consider the Hermitian curves, defined as zeros (in $\mathbb{P}^2(q)$) of the function*

$$F := X_2^s X_0 + X_2 X_0^s - X_1^{s+1}.$$

At $P_\infty := (0:0:1)$, $y := X_2/X_0$ has a pole of order $s+1$ and $x := X_1/X_0$ has a pole of order s. So $M = \begin{pmatrix} s & s+1 \\ 0 & 1 \end{pmatrix}$ describes the weighted grevlex monomial order on the monomials $x^a y^b$, with first few terms 1, x, y, x^2, yx, x^3, y^2, yx^2, x^4, y^2x, y^3. Given the function F above, these monomials are not all linearly independent. After all, any monomial involving y^k, $k \geq s$ can be reduced modulo F. (Or, if you wish, two functions with poles of the same order have a linear combination of smaller pole order.) So, for instance, for $s = 2$, $\Delta(I) := \{1, x, y, x^2, yx, x^3, yx^2\}$ is the delta set of $I := \langle F, x^4, yx^3 \rangle$. So the vector space $\langle \Delta(I) \rangle$ is a perfectly good space of (linearly independent) parity-check functions of dimension 7. There are especially many points (s^3 other than P_∞), over $GF(s^2)$ satisfying $F = 0$, so this is a good code to use over that field.

Example 2 *Consider the Klein curves, defined by the relation*

$$F := X_1^s X_2 + X_2^s X_0 + X_0^s X_1.$$

At $P_\infty := (0:0:1)$, $y := X_2/X_0$ has a pole of order s and $x := X_1/X_0$ has a pole of order $s-1$. So $M = \begin{pmatrix} s-1 & s \\ 0 & 1 \end{pmatrix}$. For $s = 3$, the monomials are ordered as above. Using x as a parity-check function precludes the use of the rational point $(0:1:0)$, since x has a pole there; but all other points of the curve are all right. So, for instance $\Delta(I) := \{1, x, y, x^2, yx, x^3, y^2\}$ is the delta set of $I := \langle x^4, yx^2, y^2x, F \rangle$. So the vector space $\langle \Delta(I) \rangle$ is a perfectly good space of (linearly independent) parity-check functions of dimension 6.

This time, there are especially many points over $GF(2^s)$ satisfying $F = 0$; so this is a good code to use over that field.

This changes drastically if one wants to use the rational point $(0:1:0)$ to get $n = 23$ instead of $n = 22$. The function x is no longer available. The functions which have poles only at P_∞ are the elements of $\mathbb{F}(q)[t_3, t_5, t_7]$ with $t_3 := y$ having weight 3, $t_5 := xy$ having weight 5, and $t_7 := x^2y$ having weight 7; and the ideal of functions relating these variables is generated by the Gröbner basis

$$\{t_5^2 + t_7 t_3,\ t_7 t_5 + t_5 + t_3^4,\ t_7^2 + t_5 t_3^3 + t_7\}.$$

(The perhaps obvious function $t_5^3 + t_3^5 + t_5 t_3$ (given that t_5^3 and t_3^5 both have the same pole order) is already an element of this ideal.)

Example 3 *Finally, consider the towers built from the Klein curves [7], [24]. These satify the functions*

$$F_i := X_i^s X_{i+1} + X_{i+1}^s X_0 + X_0^s X_i, \qquad 1 \leq i < m.$$

Now there are many (non-affine) points to be avoided if one uses the variables given. There are still, however, sufficiently many points over $GF(2^s)$ to make these interesting sources of codes.

Consider the difference in describing this tower when $m = 4$, when using the given variables and not using any of the 5 non-affine points or the origin, versus using all the points except P_∞ and changing to a new set of variables to describe all the functions which have poles only at P_∞.

Let $s = 2$. The given functions are a generating set for the ideal with common zeros the affine points (other than the origin). But a (Gröbner) basis for this is slightly larger. Let $x_1 := X_1/X_0$, $x_2 := X_2/X_0$, $x_4 := X_3/X_0$, and $x_8 := X_4/X_0$. Then the basis contains the three given functions

$$x_2 x_1^2 + x_2^2 + x_1,$$
$$x_4 x_2^2 + x_4^2 + x_2,$$
$$x_8 x_4^2 + x_8^2 + x_4;$$

and also

$$x_4^2 x_1^2 + x_4^2 x_2 + x_4 x_2 x_1 + x_1,$$
$$x_8^2 x_2^2 + x_8^2 x_4 + x_8 x_4 x_2 + x_2,$$
$$x_8^2 x_1^2 + x_8^2 x_2 + x_8 x_4 x_2 x_1 + x_8 x_1 + x_4 x_1^2 + x_4 x_2.$$

Any polynomial has a reduced form with exponent vector (a_1, a_2, a_4, a_8) satifying the property that if $a_i \geq 2$, then $a_{2i} = 0$ and $a_j < 2$ for larger j.

However, if one tries to use the other non-affine points, the variables $x_1, \ldots, x_{2^{m-1}}$ each have some poles there, so a new list of variables is needed. One choice is

$$h_8 := x_8$$
$$h_{12} := x_8 x_4$$
$$h_{14} := x_8 x_4 x_2 + x_8 x_2 x_1 + x_2$$
$$h_{15} := x_8 x_4 x_2 x_1 + x_8 x_2^2 + x_2 x_1$$
$$h_{18} := x_8^2 x_2$$
$$h_{19} := x_8^2 x_2 x_1 + x_8 x_2$$
$$h_{21} := x_8^2 x_4 x_1$$

and

$$h_{33} := h_{15} h_{18}.$$

Unfortunately, even excluding h_{33}, the list of functions needed to relate these varibles is extensive:

$$F_{12,12} := h_{12}^2 + h_8^3 + h_{12}$$

$$F_{14,12} := h_{14}h_{12} + h_{18}h_8 + h_{15}h_8 + h_{12}h_8 + h_{14}$$

$$F_{14,14} := h_{14}^2 + h_{12}h_8^2 + h_{19}$$

$$F_{15,12} := h_{15}h_{12} + h_{19}h_8 + h_8^3 + h_{12}$$

$$F_{15,14} := h_{15}h_{14} + h_{21}h_8 + h_{15}h_8 + h_{12}h_8 + h_{14}$$

$$F_{15,15} := h_{15}^2 + h_{14}h_8^2 + h_{18} + h_{15}$$

$$F_{18,12} := h_{18}h_{12} + h_{14}h_8^2 + h_{19}h_8$$

$$F_{18,14} := h_{18}h_{14} + h_8^4 + h_{21}h_8 + h_{18}h_8 + h_{15}h_8 + h_{12}h_8$$

$$F_{18,18} := h_{18}^2 + h_{12}h_8^3 + h_{14}h_8^2 + h_{19}h_8 + h_{18}$$

$$F_{19,12} := h_{19}h_{12} + h_{15}h_8^2 + f_{12}f_8^2 + f_{19}$$

$$F_{19,14} := h_{19}h_{14} + h_{18}h_{15} + h_{19}h_8 + h_8^3 + h_{21} + h_{15} + h_{12}$$

$$F_{19,15} := h_{19}h_{15} + h_{18}h_8^2 + h_{12}h_8^2 + h_{19}$$

$$F_{19,18} := h_{19}h_{18} + h_{21}h_8^2 + h_{15}h_8^2$$

$$F_{19,19} := h_{19}^2 + h_{14}h_8^3 + h_8^4 + h_{12}h_8$$

$$F_{21,12} := h_{21}h_{12} + h_{18}h_{15} + h_{14}h_8^2 + h_8^3 + h_{21} + h_{15} + h_{12}$$

$$F_{21,14} := h_{21}h_{14} + h_{19}h_8^2 + h_8^4 + h_{21}h_8 + h_{12}h_8$$

$$F_{21,15} := h_{21}h_{15} + h_{12}h_8^3 + h_{14}h_8^2 + h_{12}h_8^3$$

$$F_{21,18} := h_{21}h_{18} + h_{15}h_8^3 + h_{14}h_8^2 + h_{12}h_8^3$$

$$F_{21,19} := h_{21}h_{19} + h_8^5 + h_{21}h_8^2 + h_{18}h_8^2 + h_{15}h_8^2 + h_{14}h_8$$

$$F_{21,21} := h_{21}^2 + h_{18}h_8^3 + h_{15}h_8^3 + h_{12}h_8^3 + h_{18}h_{15} + h_{14}h_8^2 + h_{19}h_8 + h_8^3 + h_{15} + h_{12}$$

This can be simplified somewhat when working spefically over the field $GF(2^s) = GF[4]$, since the extra functions $h_i^4 + h_i$ are available. This allows us to use only four variables. But the set of functions is still rather complicated because of the presence of the origin. Without the origin, the remaining 24 rational points over $GF(4)$ are the variety gotten by using the variables h_8, h_{12}, h_{18}, and h_{14}, with functions:

$$F_1 := h_8^3 + 1$$

$$F_2 := h_{12}^2 + h_{12} + 1$$

$$F_3 := h_{18}^2 + h_{18}h_{12} + h_{18} + h_{12}$$

$$F_4 := h_{14}^2 + h_{14}h_8 + (h_{18} + 1)h_{12}h_8^2.$$

V. THE BERLEKAMP-MASSEY ALGORITHM AND THE SAKATA ALGORITHM

The syndrome matrix S mentioned above is a *Hankel* (that is, *back-shifted*) matrix. As such it can be more efficiently row-reduced (than a general matrix) by shifting early row-reductions to be used later. So, if one were to suggest to someone producing an algorithm to row-reduce a back-shifted matrix by employing shifting, but not allowing the interchange of rows, that algorithm would be akin to Berlekamp-Massey in 1-dimension, [1], [16].

In terms of matrices, what is produced, is a lower triangular matrix A such that AS (or theoretically more appropriately ASA^T) that is essentially row-reduced. This can be thought of as a change-of-basis operation from the row (and column) index functions $(1, x, \ldots)$ to new functions with these as leading monomials. The rows A_iS of AS which are zeroed out, correspond to error-locators $A_i(1, x, \ldots)^T$. In one dimension, these are all multiples of $\sigma(x)$, hence elements of the principal ideal $I := \langle\sigma(x)\rangle$. So $\sigma(x)$ corresponds to a smallest linear dependence among the rows of S.

One may also think of this by considering the 1-dimensional array of syndromes $s_0, s_1, \ldots$ as describing the syndrome matrix. The rows of A are possible linear recurrences amongst these, with σ the smallest successful such.

In N dimensions, there are still matrices S and A. S is still symmetric, but not quite back-shifted. However, if one writes out an N-dimensional array of syndromes, then every row of S can be gotten from shifting this array. So one may generalize the previous row-reduction and shifting (using the array in place of the actual matrix) to efficiently row-reduce S. This is essentially the Sakata algorithm (versions of which occur in many papers [20], [21], [22], [13], [19]); which in turn can be viewed as an N-dimensional version of Berlekamp-Massey.

The error-locators no longer form a principal ideal. What is produced is a generating set (possibly a Gröbner basis) for the error-locator ideal I, [13], as in the 1-dimensional case.

VI. FORNEY FUNCTIONS

As mentioned earlier, in one dimension, for each error position a_k (a zero of $\sigma(x)$), there is a function $\sigma_k(x)$, with $\sigma_k(x)(x-a_k) = \sigma(x)$, which is zero at all the error positions save a_k itself.

Generalizations of Forney functions have been studied in [14], [15], [8]. In N dimensions, for each error position P (a common zero of all the functions in the error-locator ideal I), there is a function $\sigma_P(\underline{x})$ with $\sigma_P(\underline{x})(x_i - x_i(P)) \in I$, which is zero at all the error positions except P. So the corresponding error magnitude can be found from

$$e_P\left(\sum_{P,f}\sigma_{P,f}f(P)\right) = \left(\sum_{P,f}\sigma_{P,f}s(f,\underline{e})\right);$$

just as in the 1-dimensional case.

References

[1] E. Berlekamp, Algebraic Coding Theory, McGraw-Hill, New York, 1968

[2] D. Cox, J. Little, D. O'Shea, *Ideals, Varieties, and Algorithms*, New York: Springer Verlag, 1992.

[3] G.-L. Feng and T. R. N. Rao, "Decoding of algebraic-geometric codes up to the designed minimum distance", IEEE Trans. on Inform. Theory, **IT-39**, pp. 37-45, 1993.

[4] G.-L. Feng, V. K. Wei, T. R. N. Rao, and K. K. Tzeng, "Simplified understanding and efficient decoding of a class of algebraic-geometric codes", IEEE Trans. on Inform. Theory, **IT-40**, pp. 981-1002, 1994.

[5] G.-L. Feng and T. R. N. Rao, "Improved Geometric Goppa Code-Part I: Basic Theory", IEEE Trans. on Inform. Theory, **IT-41**, pp. 1678-1693, 1995.

[6] G. D. Forney, "On decoding BCH codes", IEEE Trans. on Inform. Theory, **IT-10**, pp. 549-557, 1965.

[7] A. Garcia and H. Stichtenoth, "Algebraic Function Fields over Finite Fields with Many Rational Points" IEEE Trans. on Inform. Theory, **IT-41**, pp. 1548-1563, 1995

[8] J. P. Hansen, H. E. Jensen, R. Kötter, "Determination of Error Values for AG-Codes and the Forney Formula", IEEE Trans. on Inform. Theory, **IT-44**, pp. 1881-1886, 1998.

[9] T. Høholdt, J. H. van Lint, R. Pellikaan, "Order Functions and Evaluation Codes", preprint.

[10] T. Høholdt, J. H. van Lint, and R. Pellikaan, "Algebraic geometry codes", to appear in the Handbook of Coding Theory, ed. V. Pless and W. C. Huffman, Elsevier, 1998.

[11] J. Justesen, K. J. Larsen, H. E. Jensen, and T. Høholdt, "Fast decoding of codes from algebraic plane curves", IEEE Trans. on Inform. Theory, **IT-38**, pp. 111-119, 1992.

[12] T. Høholdt and R. Pellikaan, "On the Decoding of Algebraic-Goemetry Codes", IEEE Trans. on Inform. Theory, **IT-41**, pp. 1589-1614, 1995.

[13] D. A. Leonard, "Error-locator ideals for algebraic-geometric codes", IEEE Trans. on Inform. Theory, **IT-41**, pp.819-823, 1995.

[14] D. A. Leonard, "A generalized Forney formula for Algebraic-Geometric Codes", IEEE Trans. on Inform. Theory, **IT-42**, pp. 1263-1268, 1996.

[15] D. A. Leonard, "Efficient Forney functions for AG codes, IEEE Trans. on Inform. Theory, to appear

[16] J. Massey, "Shift-register synthesis and BCH decoding", IEEE Trans. on Inform. Theory, **IT-15**, pp. 122-127, 1969.

[17] M. E. O'Sullivan, "Decoding of Codes Defined by a Single Point on a Curve", IEEE Trans. on Inform. Theory, **IT-41**, pp. 1709-1719, 1995.

[18] R. Pellikaan, "On the efficient decoding of algebraic-geometric codes", Eurocode 92, edited by P.Camion, P. Charpin and S. Harari, Udine, CISM Courses and Lectures **339**, Springer-Verlag, Wien-New York, pp. 231-253, 1993.

[19] K. Saints and C. Heegard, "Algebraic-Geometric Codes and Multidimensional Cyclic Codes: A Unified Theory and Algorithms for Decoding Using Gröbenr Bases", IEEE Trans. on Inform. Theory, **IT-41**, pp. 1752-1768, 1995.

[20] S. Sakata, "Finding a minimal set of linear recurring relations capable of generating a two-dimensional array", Journal of Symbolic Computation **5**, pp. 321-327, 1988.

[21] S. Sakata, "Extension of Berlekamp-Massey algorithm to N dimensions", Information and Computation **84**, pp. 207-239, 1990.

[22] S. Sakata, J. Justesen, Y. Madelung, H. E. Jensen, and T. Høholdt, "Fast decoding of AG-codes up to the designed minimum distance", IEEE Trans. on Inform. Theory, **IT-41** , pp. 1672-1677, 1995.

[23] S. Sakata, D. A. Leonard, H. E. Jensen, T. Høholdt, " Fast Erasure and Error Decoding of Algebraic Geometry Codes up to the Feng-Rao Bound", IEEE Trans. on Inform. Theory, **IT-43**, to appear, July, 1998.

[24] C. Voss and T. Høholdt, "An Explicit Construction of a Sequence of Codes Attaining the Tsfasman-Vlăduţ -Zink Bound: The First Steps" IEEE Trans. on Inform. Theory, **IT-43**, pp.128-135, 1997.

On the Coset Weight Distributions of Z_4-Linear Goethals Codes and Some Relations to Kloosterman Sums*

Tor Helleseth
Department of Informatics
University of Bergen
Høyteknologisenteret,
N-5020 Bergen
Norway
torh@ii.uib.no

Victor Zinoviev
Institute for Problems of Inform. Transm.
Russian Academy of Sciences
Bol'shoi Karetnyi per. 19
Moscow 101447
Russia
zinov@mail.iipt.ru

Abstract

We study the coset weight distribution of the two classes of codes: Z_4 linear Goethals codes (which we denote G_4), over Z_4 with Lee minimum distance 8, of length $n = 2^m$ (m odd) and binary Z_4 linear Goethals codes (which we denote by G_2) with Hamming distance 8, of length $N = 2n = 2^{m+1}$. We show some connections between the weight distributions of these Goethals codes and Kloosterman sums. In particular we show that the number of distinct weight distributions for cosets of weight four in the Goethals codes, increases with the length. We also obtain new results on Kloosterman sums from the coset weight distributions of the Goethals codes.

1 Introduction

This paper is initiated by the papers of Charpin-Zinoviev [3] and Helleseth-Kumar [6]. We examine here the coset weight distributions of the two classes of codes: Z_4-linear Goethals codes (which we denote G_4) over Z_4 with Lee minimum distance 8, of length $n = 2^m$ (m odd) and binary Z_4-linear Goethals codes (which we denote G_2) with Hamming minimum distance 8, of length $N = 2n = 2^{m+1}$. The main result of [3] is: *the number of weight distributions of cosets of* 3-*error-correcting binary BCH-codes increases with the value of* m. Therefore it gives an example of an infinite class of codes whose external distance (i.e. the number of different nonzero values, which occur as the weight of code words in the dual code) is constant while the number of distinct coset weight distributions increases with the length. We prove here that for both of the Z_4-linear Goethals codes we have the same phenomena: *the number of weight distributions of cosets of the binary Z_4-linear (or, Z_4-linear over Z_4) Goethals code increases with the value of* m. We give the exact expressions for the number of cosets of weight 4, with a given number of code words of weight 4, in terms of the Kloosterman sums. We essentially use here the results on uniformly packed codes [1], [2] and [4], and the results and the approach used in [3] and [6].

---

*This research is supported in part by The Norwegian Research Council.

A linear code $\mathcal{C}$ over Z_4 with length n is an additive subgroup of Z_4^n. The Gray map $\phi : Z_4 \to Z_2^2$ is defined by $\phi(0) = 00$, $\phi(1) = 01$, $\phi(2) = 11$, and $\phi(3) = 10$. Hammons, Kumar, Calderbank, Sloane and Solé [5] have shown that efficient nonlinear codes such as Kerdock, Preparata, etc., can be very simply constructed as binary images under the Gray map of linear codes over Z_4.

Let R_m be a Galois ring of characteristic 4 with 4^m elements and let R_m^* be the set of units of R_m. This set R_m^* has a multiplicative cyclic subgroup of order $2^m - 1$. Let $\mathcal{T}_m = \{0, 1, \beta, ..., \beta^{2^m-2}\}$, where $\beta \in R_m^*$ is an element of order $2^m - 1$.

The Z_4-linear Goethals code G_4 is the code over Z_4 with parity-check matrix H_G given by

$$H_G = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & \beta & \beta^2 & \cdots & \beta^{2^m-2} \\ 0 & 2 & 2\beta^3 & 2\beta^6 & \cdots & 2\beta^{3(2^m-2)} \end{bmatrix} . \qquad (1)$$

In Hammons, Kumar, Calderbank, Sloane and Solé [5], it is shown that if m is odd, then G_4 has minimum Lee weight 8 and its Gray map which we denote $G_2 = \phi(G_4)$ gives a $(2^{m+1}, 2^{2^{m+1}-3m-2}, 8)$ binary nonlinear code which has the same weight distribution as the Goethals code. Denote such arbitrary Goethals-like code by G.

Here we give some results about the coset weight distribution of any code G, and the coset weight distribution of the code G_2 (and, therefore, of G_4).

Let C be any binary code of length n and let ρ be its covering radius. We say that such a code is *uniformly packed*, in the sense of [1], if there exist rational numbers $\alpha_0, \ldots, \alpha_\rho$ such that for any $v \in GF(2)^n$

$$\sum_{k=0}^{\rho} \alpha_k \; f_k(v) = 1 , \qquad (2)$$

where $f_k(v)$ is the number of codewords at distance k from v. Any Goethals-like code G of length N is uniformly packed and has:

$$\begin{array}{ll} \alpha_0 = \alpha_1 = 1, & \alpha_2 = (19N - 34)/2N(N-1), \\ \alpha_3 = 15/2(N-1), & \alpha_4 = 60/N(N-1), \\ \alpha_5 = 30/(N-1)(N-4), & \alpha_6 = 180/N(N-1)(N-4) . \end{array} \qquad (3)$$

As it follows from the result of [5], there is a full similarity between the coset weight distributions of the binary Z_4-linear Goethals-like codes G_2 of length $N = 2^{m+1}$ and the Z_4-linear Goethals codes G_4 of length $n = 2^m$ over Z_4.

Let D be a coset of G. The weight of the coset D is the minimum weight of the codewords of D. A leader of D is a codeword of D of minimum weight. A coset D is called an orphan, if the the union of the support of all the leader of D covers the coordinate set of D. Denote by $\mu_i = (\mu_{i,0}, \mu_{i,1}, ..., \mu_{i,n})$ the weight distribution of D. As G is uniformly packed in the sense of [1], to find all vector μ_i, it is enough to find the values $\mu_{i,j}$ for $j = 0, 1, ..., 5$.

Proposition 1 *Let G be any Goethals-like code.* (i) *There is only one coset weight distribution for the cosets of G of weight* 1, 2 *and* 3. *The number of codewords of weight* 6 *in the coset of weight* 2 *is*

$$\mu_{2,6} = \frac{1 - \alpha_2}{\alpha_6} = \frac{(N-4)(2N^2 - 21N + 34)}{360} .$$

The number of codewords of weight 5 *in the coset of weight* 3 *is:*

$$\mu_{3,5} = \frac{1-\alpha_3}{\alpha_5} = \frac{(N-4)(2N-17)}{60}.$$

(ii) *There are* $\Gamma_2(5) = N(N-2)(N+8)/12$ *distinct cosets of* G *of weight* 5. *All of these cosets have the same weight distribution and each of them contains* $\mu_{5,5} = (N-1)(N-4)/30$ *codewords of weight* 5.

Proposition 2 (i) *For the code* G_2 *of length* N *there are* $(N-1)(N-2)(3N+2)/12$ *distinct cosets of weight* 4. *Let* D *be any coset of weight* 4 *of* G_2. *The weight distribution of* D *is uniquely defined by the value* $\mu_{4,4}$. *The number of codewords of weight* 6 *in* D *is:*

$$\mu_{4,6} = \frac{1-\alpha_4\mu_{4,4}}{\alpha_6} = \frac{N(N-1)(N-4)}{180} - \frac{1}{3}(N-4)\mu_{4,4}.$$

(ii) *For the code* G_2 *of length* N *there are* $(N-2)(N+8)/12$ *distinct cosets of weight* 6. *All of these cosets have the same weight distribution. Such a coset is an orphan and it contains* $\mu_{6,6} = N(N-1)(N-4)/180$ *codewords of weight* 6.

We know all orphans of G_2 of weight 4 and we can express the number $\mu_{4,4}$ for any coset of weight 4 with a given syndrome in terms of Kloosterman sums. Here we give only a few typical results. Let $Tr(x)$ denote the trace function from $F = GF(2^m)$ to $GF(2)$. The well known Kloosterman sums are defined for each $a \in F^* = F \setminus \{0\}$ by

$$K(a) = \sum_{x \in F^*} (-1)^{Tr(ax+\frac{1}{x})}.$$

We need the following result due to Lachaud and Wolfmann [10].

Lemma 1 *The set of the* $K(a)$, $a \in F^*$ *is the set of all the integers* $s \equiv -1 \pmod 4$ *in the range* $[-2^{(m/2)+1},\ 2^{(m/2)+1}]$.

Consider an error with locations X, Y, Z and U with $e_X = e_Y = e_Z = 1$ and $e_U = -1$. This leads to the equations over $\mathcal{T}$

$$\begin{array}{rcl} X + Y + Z - U &=& A + 2B \\ 2X^3 + 2Y^3 + 2Z^3 - 2U^3 &=& 2C \end{array}$$

which is equivalent to the equations over F

$$\left.\begin{array}{rcl} x + y + z + u + &=& a \\ u^2 + xy + xz + xu + yz + yu + zu &=& b^2 \\ x^3 + y^3 + z^3 + u^3 &=& c\,. \end{array}\right\}$$

Proposition 3 *Let* $a, b,$ *and* c *be any fixed elements of* F, *where* $a \neq 0$. *The number of codewords of weight four in the coset* D *defined by the system above is equal to the number* $M(b,c)$, *where*

$$\begin{array}{rcl} M(b,c) &=& \frac{1}{3} \times 2^{m-1} + \frac{1}{6}((-1)^{Tr(k_1)}K(k_1k_2) - 3 \\ &+& (-1)^{Tr(b)}(2-(-1)^{Tr(k_1)}-(-1)^{Tr(k_2)}))\,, \end{array}$$

where

$$k_1 = b^2 + c + 1 \quad \text{and} \quad k_2 = b^2 + b + c + \sqrt{c},$$

The number of different such coset, when $a, b,$ *and* c *run over* F *where* $a \neq 0$, *is equal to*

$$\Gamma(2,a,b,c) = 2(2^m-1)^2(2^m-2).$$

The connection above to Kloosterman sums combined the result of Lachaud and Wolfmann gives a natural explanation the phenomena, mentioned in [3] for 3-error-correcting BCH codes of length 2^m (m odd): that the number of distinct coset weight distributions increases with the length.

By calculating the number of words of weight four in some special cosets of the Goethals codes we can express these numbers in terms of Kloosterman sums. In particular since these numbers must be integers a detailed study will lead to conditions on Kloosterman sums. The following results on Kloosterman sums were obtained in this way [9].

Proposition 4 *Let m be any positive integer $m \geq 3$. Let a be an arbitrary element of F^*. Then if $Tr(a) = 1$*

$$K(a) \equiv 3 \pmod 8,$$

and if $Tr(a) = 0$

$$K(a) \equiv 7 \pmod 8.$$

Proposition 5 *Let m be odd and a any element of F^*, then*

$$K(a^4 + a^3) \equiv 3 \pmod{12}.$$

Let m be even, and a be any element of F^*, then

$$K(a^4 + a^3) \equiv \begin{cases} 7 & \pmod{12}, \quad \text{if } Tr(a) = 0, \\ 11 & \pmod{12}, \quad \text{if } Tr(a) = 1. \end{cases}$$

Detailed results on the coset weight distributions of Goethals-like codes G_2 and G_4 will be published in [8] and the relations to Kloosterman sums will be published in [9].

References

[1] L.A. Bassalygo, G.V. Zaitsev and V.A. Zinoviev, "Uniformly packed codes", *Problems Inform. Transmiss.*, vol. 10, No 1, pp. 9-14, 1974,

[2] L.A. Bassalygo and V.A. Zinoviev, "Remark on uniformly packed codes", *Problems Inform. Transmiss.*, vol. 13, No 3. pp. 22-25, 1977,

[3] Charpin and V.A. Zinoviev, "On coset weight distributions of the 3-error-correcting BCH codes", *SIAM J. of Discrete Math.*, vol. 10, No. 1, pp. 128-145, February 1997.

[4] J.M. Goethals and H.C.A. Van Tilborg, "Uniformly packed codes", *Philips Res. Repts*, vol. 30, 9-36, 1975.

[5] A. R. Hammons, P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Sole, "The Z_4-linearity of Kerdock, Preparata, Goethals, and related codes," *IEEE Trans. Inform. Theory*, vol. 40, pp. 301-319, Mar. 1994.

[6] T. Helleseth and P. V. Kumar, "The algebraic decoding of the Z_4-linear Goethals code," *IEEE Trans. Inform. Theory*, vol. 41, pp. 2040-2048, Nov. 1995.

[7] T. Helleseth, P. V. Kumar, and A. Shanbhag, "New codes with the same weight distributions as the Goethals codes and the Delsarte-Goethals codes," *Designs, Codes and Cryptography*, vol. 9, No. 2, pp. 257-266, 1996.

[8] T. Helleseth and V.A. Zinoviev, "On Coset Weight Distribution of the Z_4-Linear Goethals Code," under submission.

[9] T. Helleseth and V.A. Zinoviev, "On Z_4-Linear Goethals Codes and Kloosterman sums," under submission.

[10] G. Lachaud & J. Wolfmann, "The Weights of the Orthogonals of the Extended Quadratic Binary Goppa Codes", *IEEE Trans. on Inform. Theory*, vol. 36, No. 3, pp. 686-692, May 1990.

Further Results on a Conjectured 2-level Autocorrelation Sequence

Anchung Chang
anchungc@milly.usc.edu

Tor Helleseth*
torh@ii.uib.no

P. Vijay Kumar†
kumar@aditya.usc.edu

Abstract

Recently, No, Golomb, Gong, Lee and Gaal conjectured that a certain family of binary sequences possesses ideal 2-level autocorrelation function. Subsequently, Chang, Gaal, Golomb, Gong and Kumar showed that if C is the extended linear, cyclic code generated by the sequence and its cyclic shifts, then (i) the sequence has the balance property, (ii) the weight of each codeword in C is divisible by 2^m where $q = 2^{2m+1}$ is the length of the code and (iii) the dual $C^\perp$ has minimum distance at least equal to 7.

In this paper, an overview of these developments will be provided as well as a recent simple observation that shows that C has at most 5 distinct nonzero weights. Furthermore C has the same 5-level weight distribution as the dual of the triple-error correcting BCH code.

1 Overview

For any integer $k \geq 1$, let $\mathbf{F}_{2^k}$ denote the finite field of 2^k elements. Let $m \geq 2$ be an integer and $n = 2m+1$. Let $T : \mathbf{F}_{2^n} \to \mathbf{F}_2$ denote the trace function given by

$$T(x) = \sum_{i=0}^{n-1} x^{2^i}, \quad x \in \mathbf{F}_{2^n} .$$

Set α be a primitive element of $\mathbf{F}_{2^n}$ and set $r = 2^{m+1}+1$. No, Golomb, Gong, Lee and Gaal [1] conjectured that the sequence

$$s(t) = T(\alpha^t + \alpha^{rt} + \alpha^{r^2 t})$$

has the ideal autocorrelation function, i.e.,

$$\sum_{t=0}^{2^n-2} (-1)^{s(t+\tau)+s(t)} = \begin{cases} 2^n - 1 & \text{if } \tau = 0 \\ -1 & \text{else} . \end{cases}$$

---

[1] This work was supported in part by the National Science Foundation under Grant Number NCR-9612864 and in part by Norwegian Research Council.

[*] Tor Helleseth is with the Department of Informatics, University of Bergen, Høyteknologisenteret, N-5020 Bergen, Norway.

[†] Anchung Chang and P. Vijay Kumar are with the Communication Sciences Institute, EE-Systems, EEB534, University of Southern California, Los Angeles, CA 90089-2565, USA.

A more recent paper of Dobbertin [2] presents a family of sequences with conjectured ideal autocorrelation function that includes this sequence as a special case.

Consider the linear cyclic code

$$C = \{\, Tr(ax + bx^r + cx^{r^2})\ x \in \mathbf{F}_{2^n} \mid a, b, c \in \mathbf{F}_{2^n}\}.$$

Chang, Gaal, Golomb, Gong and Kumar [3] showed the following:

Theorem 1 *$s(t)$ has the balance property.*

Theorem 2 *The minimum distance of the dual of C is ≥ 7.*

Theorem 3 *The Hamming weight of each codeword in C is divisible by 2^m.*

and conjectured that C has at most 5 distinct nonzero weights.

An outline of the proof of this conjecture will now be given.

Consider the triple-error correcting BCH code having zeros $\{1, 3, 5\}$

$$C' = \{\, Tr(ax + bx^3 + cx^5)\ x \in \mathbf{F}_{2^n} \mid a, b, c \in \mathbf{F}_{2^n}\}.$$

The minimum distance of the duals of both C and C' is at least 7. It then follows from the Pless power-moment identities [4] [5] [6] that the 2nd, 4th and 6th moments of C and C' are the same.

From Theorem 3, the Hamming weight of each codeword in C is divisible by 2^m. By proceeding them along the lines of the proof of Lemma 3 of Canteaut, Charpin and Dobbertin [7], it can be showed that C has at most 5 distinct nonzero weights.

References

[1] J-S. No, S. W. Golomb, G. Gong, H-K. Lee and P. Gaal, "New Binary Pseudorandom Sequences of Period $2^n - 1$ with Ideal Autocorrelation', *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 814-817, March 1998.

[2] H. Dobbertin, "Kasami Power Functions, Permutation Polynomials and Cyclic Difference Sets" NATO-A.S.I. Workshop, Aug. 1998, Bad Windsheim, Germany.

[3] A. Chang, P. Gaal, S. W. Golomb, G. Gong and P. V. Kumar, "On a Sequence Conjectured to Have Ideal 2-level Autocorrelation Function", ISIT Aug. 1998, MIT, Cambridge, MA, USA.

[4] P. V. Kumar and C-M. Liu, "On Lower Bounds to the Maximum Correlation of Complex Roots-of-Unity Sequences" *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 633-640, May 1990.

[5] F. J. McWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes.* Amsterdam, The Netherlands: North Holland, 1979.

[6] T. Kasami, "Weight Distributions of Bose-Choudhary-Hochquengham Codes", in *Combinatorial Mathematics and Its Applications.* Chapel Hill, NC: Univ. North Carolina Press, 1969.

[7] A. Canteaut, P. Charpin and H. Dobbertin, "Binary m-Sequences with Three-Valued Crosscorrelation: A Proof of Welch's Conjecture, " preprint, 1998.

Controller Fragility: What's All The Fuss?

David Faris[*] Thomas Paré[†]

Andrew Packard[‡] Khalid Al Ali[§] and Jonathan P. How[¶]

Abstract

This paper is intended primarily as a rebuttal to *"Robust, Fragile, or Optimal?"* [KB97], which quantifies a certain measure of the sensitivity of closed-loop stability to perturbations in the controller. We present analysis techniques which clearly indicate that the controller uncertainty models and robust stability metric defined in [KB97] are physically unrealistic and typically pessimistic. As a result the conclusions in [KB97] are misleading, and are not consistent with control engineering practice and experience. Our current focus is on developing more relevant analysis approaches for a given plant model/set and a given controller. Detailed physical (computational) uncertainty models are proposed for continuous-time (discrete-time) controllers. Well established tools taken directly from modern robust control were used to compute allowable uncertainties such that various closed-loop objectives are satisfied. Several recently cited examples are presented in detail. The results clearly show that, in most cases, the controller/closed-loop systems that have been labeled *"fragile"* are in fact quite robust to controller perturbations.

1 Introduction

This paper is intended primarily as a rebuttal to *"Robust, Fragile, or Optimal ?"*[KB97], and some of the ideas put forth at session **FA09** at the 1998 American Control Conference [Dor98]. Ref. [KB97] attempts to quantify a certain measure of the closed-loop sensitivity to controller perturbations, and clearly requires some kind of measured response, especially with regard to how the conclusions in the paper impact current practices in applied robust control. We will concentrate on clarifying some issues regarding sufficiency conditions for closed-loop robustness and the associated controller uncertainty models.

The uncertainty modeling and analysis tool applied in [KB97] is the *"Parametric Stability Margin"* defined in [BCK95]. Specifically, the coefficients of the nominal controller's transfer function are modeled with independent, interval bounded uncertainty.

---

[*] dwfaris@me.berkeley.edu, Dept. of Mech. Engr., Univ. of California - Berkeley, Berkeley CA 94720

[†] tpare@sun-valley.stanford.edu Dept. of Mechanical Engr, Stanford University, Stanford CA 94305

[‡] packard@me.berkeley.edu, Department of Mechanical Engineering, University of California - Berkeley

[§] khalid@me.berkeley.edu, Department of Mechanical Engineering, University of California - Berkeley

[¶] howjo@sun-valley.stanford.edu Department of Aeronautics and Astronautics, Stanford University

Some measure of the smallest distance in the coefficient space to a region of closed-loop instability is then computed. While we do not dispute any of the computations regarding this instability measure, we find the results very misleading. We believe that there are other deterministic measures of sensitivity that are simultaneously more relevant and more optimistic. Moreover, these conclusions are supported with Monte Carlo simulations of various controller implementations. These combined results suggest that, for the most part, the controllers labeled *fragile* in [KB97] are quite robust to implementation errors.

The modern approach to robustness analysis of feedback control systems (largely credited to Safonov and Doyle, with many of the earliest ideas first presented at Allerton) directly addresses a detailed, component level description of a feedback system. Each and every aspect of the dynamic system model is assessed, and the potential uncertainty (or variation) is categorized. Depending on the types of assumptions made about the uncertainties, some sort of worst-case analysis is performed to verify that the dynamic system has certain desirable properties for all possible component variations. Most robustness analysis research since 1978 has focused on theoretical and computational issues for these worst-case analysis tests. This modern viewpoint should be contrasted with earlier robustness methods, such as gain and phase margins, and robustness to "percentage" variations in the plant. These methods are characterized by simple computational or graphical tests, and the crude information they yield. Very little explicit information is obtained regarding the impact that variations in specific physical parameters have on the closed-loop system.

Certainly, if one wants to test the sensitivity of the closed-loop system to variations that arise due to our inability to exactly implement a given controller, then the details of how the controller is implemented and typical uncertainties associated with any given implementation must be carefully considered, and the modern approach should be used. Different implementations of the same controller may demonstrate vastly different behaviors to perturbations in the physical parameters of which they are composed. Indeed, our terminology, at times, may appear overly physical, making reference to filters, op-amps, resistors, capacitors, *etc.* However, the systems and circuits literature dating back almost 30 years is resplendent with excellent examples of this methodology of sensitivity analysis, see Refs. [Par71, RG71, EV81, Mos70, SE75, GB81, SM71]. We maintain that the work in this paper is a natural evolution of this uncertainty & sensitivity modeling process.

Depending on the design objectives, it is also possible that the closed-loop system is highly sensitive to the controller, as an operator. In this case, it is likely that the closed-loop system will be sensitive to any implementation of the controller. There is a need to distinguish and separate this idea from the above discussion. A useful starting point might be to make the following distinction:

- **Robustness of the Closed-loop System to Perturbations in the Controller**, referring to the point of view of treating the controller as a single entity – a linear operator with uncertainty. We will not focus on these issues, but the robustness properties could be characterized in terms of the largest gap metric ball around the nominal controller for which adequate closed-loop performance is maintained, [GG94], [YP95].

- **Fragility of the Controller Realization**, referring not to a controller in general, but to a specific implementation of a controller and the associated uncertainty model of this implementation. The focus should be on the inherent robustness of the closed-loop system to perturbations in the physical parameters which make up the implementation.

Note that some special synthesis problems, such as optimal gain margin, essentially have closed-form solutions. Nevertheless, these limited objectives are not appropriate for practical control design, as only one limited objective (such as gain margin) is included in the cost function. These nice theoretical results should be used to establish bounds on achievable properties (which may help assess constrained sub-optimal behavior) and/or as introductory teaching tools. However, these optimization problems are of little practical interest as it is well known that they often result in controllers that have very small robustness (gain and/or phase) margins. As a result, the closed-loop system is then typically very sensitive to the controller as an operator. Perhaps this point needs to be made more strongly in some of the cited examples (such as [DFT92]), where many of these specialized techniques are posed and solved. The distinction between the two cases given above will play a central role in assessing the controller fragility in the examples in Ref. [KB97].

The paper these ideas in more detail and analyzes the fragility of several controllers using a variety of techniques. Section 2 outlines the realization, uncertainty modeling, and robustness analysis techniques used to re-analyze and reinterpret the examples from Ref. [KB97]. Section 3 then gives a short summary of these results, and Section 4 provides concluding remarks.

2 Methodology

2.1 Background

Our procedure for assessing fragility has two main points of emphasis: (1) physically motivated parametric uncertainty models of the implemented controller; (2) computational methods to analyze the general closed-loop robustness problem incorporating both plant and controller uncertainty models. Details for continuous-time and discrete-time systems are presented in Sections (2.2) and (2.3), respectively. Currently, only LTI-SISO systems are considered. However, the general methodology and uncertainty models should be applicable to LTI-MIMO controllers, with controller realization based on the bi-diagonal state-space transformation.

2.2 Continuous-Time Controllers

This section covers the details of our implementation model for continuous-time controllers, which is based on reasonably simple physical models whose uncertain/dynamic behavior is easily derivable or well established. More specifically, continuous-time controllers are implemented as cascaded interconnections of first- and second-order **Active-RC Filters**. Each filter is composed of distinct physical elements (such as op-amps, resistors, and capacitors). Thus, the uncertainty model obtained for any given controller is a function of: (1) the topology of the interconnection in general terms; (2) the specific design of each filter in the interconnection; and (3) real uncertainty in actual physical parameters (hereafter referred to as *parametric uncertainty*). Given some nominal continuous-time controller transfer function, $C_0(s)$, we proceed as follows:

Step CT-1: **Expand Nominal Controller.** The nominal controller is expressed as a cascade of real first- and second-order filters. The pairing of poles and zeros for minimum sensitivity is still an open issue. Our initial designs were based on choices which produced filters with the flattest possible dynamic response. See [Mos70] for an alternative point of view and other references on this topic. Note that we only use second order filters to implement complex-conjugate transfer function poles or zeros.

Step CT-2: **Design Active-RC Filters.** Each term of the controller's expansion, as well as any additional elements necessary for its interconnection, is designed as an Active-RC filter. Figure (1) shows one of our four basic filters used for this purpose. The differential/algebraic equations for each filter are expressed as LFT's (Linear Fractional Transformations) on the set of parametric uncertainties for each resistor/capacitor therein. To specify resistor and capacitor values for any given filter, we either solve the underdetermined system of equations or use standard filter design tools, which may apply additional optimization criteria.

Step CT-3: **Compose Controller Uncertainty Model.** The controller is implemented as an interconnection of the filters designed in the previous step. As a result of the fact that each filter is modeled as an LFT on the set of real parametric uncertainties, the resulting controller uncertainty model is an LFT on the union of the set of real parametric uncertainties in each filter.

2.3 Discrete-Time Controllers

The treatment of continuous-time controllers in Section (2.2) produces uncertainty models with real physical parameter dependence and natural parametric uncertainty. Considering now discrete-time controllers, the logical question to ask is: *What are possible sources of uncertainty in discrete-time filters, and how might they be incorporated in an uncertainty model ?* Our approach in this regard is to consider the effects of **Floating-Point Errors** [1]. While there are possibly other sources of uncertainty in discrete-time filters not addressed by this approach, we believe that floating-point representation errors and floating-point operation errors are logical starting points.

Our treatment of floating point arithmetic is standard (see, for example, [GVL96] pp. 59-69). Let N denote the set of integers, R the set of real numbers, and F the set of floating point numbers, with which we associate a floating point base, $\beta \in \mathrm{N}_+$. and the number of place holders, $t \in \mathrm{N}_+$. Let the pair $\{(L, U) \mid L \in \mathrm{N}, U \in \mathrm{N}, L \leq U\}$ denote the lower and upper exponent limits, respectively. The set of floating point numbers for

---

[1] **Note:** Throughout this section, *"t"* is used to denote the number of floating point place holders. At no point in this section should *"t"* be interpreted as the independent variable *"time"*.

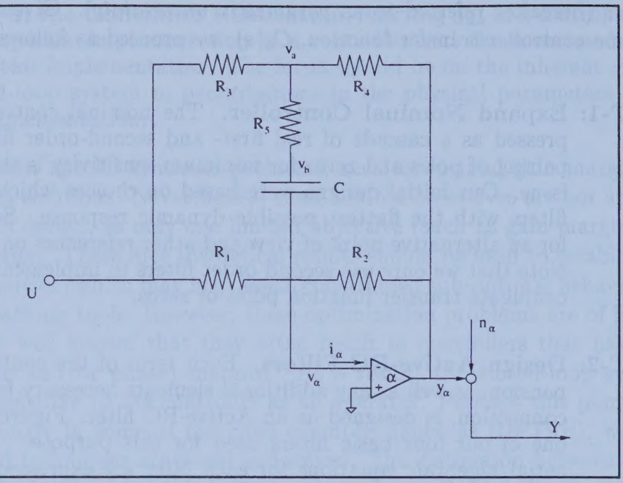


Figure 1: Active-RC Single-Amplifier, First-Order, Inverting, Minimum-Phase Filter

the above system of representation is:

$$\mathbf{F}(\beta,t,L,U) \;:=\; \left\{\beta^e \cdot \sum_{i=1}^{t} d_i\beta^{-i} \;\middle|\; d_i \in \mathbf{N}_{[0,\beta)},\ e \in \mathbf{N}_{[L,U]}\right\} \;\cup\; \left\{0\right\} \tag{1}$$

To represent a floating point number using Eq. (1), we assume that L and U are unconstrained and that the modeling computations are performed at significantly higher precision than is implemented in the floating point set. This provides the basis for representing constants that appear in the discrete-time filters. To represent uncertainty in performing discrete-time computations, consider the set of basic floating point operations:

$$\{(\mathrm{x}\ op\ \mathrm{y}) \mid \mathrm{x} \in \mathbf{R}, \mathrm{y} \in \mathbf{R}, op \in \{+,-,\times,\div\}\}$$

The floating point result of the basic operations above is known to satisfy Equations (2)–(3), with $\mathcal{U}_e$ denoting the so-called *Unit Floating Point Round-Off Error*:

$$fl(\mathrm{x}\ op\ \mathrm{y}) \;:=\; (\mathrm{x}\ op\ \mathrm{y})\,(1 + \mathcal{U}_e\delta) \tag{2}$$

$$op \;\in\; \{+,-,\times,\div\}\ ,\quad \mathcal{U}_e \;:=\; \frac{1}{2}\beta^{1-t}\ ,\quad \delta \;\in\; \mathbf{R}_{[-1,1]} \tag{3}$$

Therefore, given some nominal discrete-time controller transfer function, $C_0(z)$, we proceed as follows:

Step DT-1: Expand Nominal Controller. See Step CT-1.

Step DT-2: Implement Discrete Time Filters. First-Order and Second-Order Filters are written in controllable canonical state-space form, as shown in Sections (2.3.1) and (2.3.2). All constants are submitted for floating point representation. The appropriate floating point perturbation behavior is then obtained as an LFT.

Step DT-3: Compose Controller Uncertainty Model. See Step CT-3.

2.3.1 Discrete-Time First-Order Filter

The transfer function and controllable canonical form for a discrete-time first-order filter are:

$$G(z) = \frac{Y(z)}{U(z)} = \frac{b_0 z + b_1}{z + a_1} \tag{4}$$

$$\begin{bmatrix} x(k+1) \\ y(k) \end{bmatrix} = \left[\begin{array}{c|c} -a_1 & 1 \\ \hline b_1 - a_1 b_0 & b_0 \end{array}\right] \begin{bmatrix} x(k) \\ u(k) \end{bmatrix} \tag{5}$$

The actual floating point result for the controllable canonical form is shown in Equations (6)–(7), with $\left\{\hat{A}, \hat{C}, \hat{D}\right\}$ denoting the floating point representations of $\{-a_1, b_1 - a_1 b_0, b_0\}$:

$$x(k+1) = \left[\hat{A}\, x(k) \left(1 + \mathcal{U}_e \delta_1(k)\right) + u(k)\right]\left(1 + \mathcal{U}_e \delta_2(k)\right) \tag{6}$$

$$y(k) = \left[\hat{C}\, x(k) \left(1 + \mathcal{U}_e \delta_3(k)\right) + \hat{D}\, u(k) \left(1 + \mathcal{U}_e \delta_4(k)\right)\right]\left(1 + \mathcal{U}_e \delta_5(k)\right) \tag{7}$$

The LFT is specified on the set of five scalar perturbations, $\{\delta_1(k), \cdots, \delta_5(k)\}$, which are bounded by $[-1, 1]$ and are time varying to account for the fact that each floating point operation is subject to a different perturbation (bounded by $\mathcal{U}_e$) depending on the current arguments.

2.3.2 Discrete-Time Second-Order Filter

The discrete-time second-order filter is treated in the same manner, resulting in 9 time-varying perturbations. This will be included in the final paper.

2.4 Analysis Methods

Once the uncertainty models have been constructed, it is a straightforward exercise to apply standard results from robust control to conservatively determine the allowable range of controller parameter variations for which closed-loop stability is maintained. In the Continuous-Time examples, we treat the uncertainty in the resistors and capacitors as real, time-invariant parametric uncertainty, and use the structured singular value upper bound, [FTD91], on a densely gridded frequency response. In the Discrete-Time examples, we treat the δ_i parameters as real, time-varying, and apply an optimally-scaled small-gain theorem to conservatively determine stability of the closed-loop discrete-time system. Both of these tests give guaranteed robust stability answers to the perturbation models. As an alternative, we also make statistical calculations, where the uncertain parameters (resistors and capacitors) are given probability distributions, and the probability of closed-loop stability is approximated with a Monte-Carlo calculation, as described in [RS93]. All these methods described above are invariant to unit-changes and time-scaling, which is desirable, as only the intrinsic uncertainty in the components should effect the robustness results.

3 Results

We analyzed the various examples cited in [KB97] using the approach in Section 2. The deterministic results are summarized in Table 1. The column headed "PSM" (parametric stability margin) summarizes the results from [KB97]. The column headed "RC" shows the results from the new analysis presented in this paper. In particular it shows the allowable componentry uncertainty percentages for which closed-loop stability is guaranteed. The column headed "Bits" shows the value of t (with $\beta = 2$) for which the floating point representation of the discrete controller is guaranteed to be stabilizing.

For all designs considered in the analysis, stable closed loop systems could be achieved using realizations well within the current state-of-the-art implementations. For instance, an 8-bit realization of the discrete controller for design example #5 is guaranteed to give a stable closed loop system. Guaranteed stability using only single precision seems to contradict the fragility of the design as assessed by the parametric stability margin. In this case, with the PSM measure less than 1 part in 10,000, the authors in [KB97] claim that this design is too fragile for implementation. However, our analysis indicates this system would be easy to implement and would be stable with DSP components that became common in the late 1970's.

However, perhaps the most striking contradiction between the PSM and standard engineering practice occurs in design example #4, which is a μ–control design for a magnetic suspension system, based on the original work of [FNMU95]. There, an electromagnetic suspension system is modeled, and a μ-Synthesis approach is taken to design a robust performance controller. The nominal plant is given in Equation (8), while the plant uncertainty set is modeled with additive uncertainty weight given in Equation (9). The feedback interconnection is a standard unity gain, negative feedback regulation problem, with performance weight across plant output disturbance to plant output error. The performance weight is given in Eq. (9).

$$G_{\mathrm{nom}}(s) := \frac{-36.27}{(s+66.94)\,(s-66.94)\,(s+45.69)} \tag{8}$$

$$W_u(s) := \left(9.7794 \times 10^{-7}\right) \frac{(s+8)(s+170)(s+420)}{(s+30)(s+35)(s+38)}, \quad W_p(s) := \frac{20}{s+0.1} \tag{9}$$

μ-Synthesis is a two-step iterative process. Two controllers are presented in the original paper; these are denoted by $K_1(s)$ in Equation (4.12) and $K_2(s)$ in Equation (4.14) of the original citation. They will not be repeated here. They are both stable, with real left-half-plane zeros, and several second-order poles (i.e: complex-conjugate poles). The authors of [KB97] chose to analyze controller $K_1(s)$. Both controllers, $K_1(s)$ and $K_2(s)$, were realized according to the methods described in Section (2). However, for brevity, only the realization associated with controller $K_1(s)$ is presented, in Equation (10) below. The two realizations are very similar, with K_2 having one additional second order filter in its cascade. The controller realization for $K_1(s)$ has a total component count of 47 (7 capacitors), while $K_2(s)$ has 63 (9 capacitors).

$$\begin{aligned} K_1(s) = {} & \left[-5220 * \frac{s+66.94}{s+734.70}\right]\left[-1000 * \frac{s+12.00}{s+39.77}\right] \\ & \left[+1000 * \frac{s^2+83.69s+1736.22}{s^2+630.40s+207987.20}\right]\left[+1 * \frac{s^2+65.00s+1050.00}{s^2+63.20s+1024.77}\right] \\ & \left[-0.10 * \frac{1}{s+0.10}\right] \end{aligned} \tag{10}$$

In this case, the PSM for the controller is 1.5×10^{-15} and is thus considered **very** fragile. Using the combined design/analysis method described in Section 2, it was found

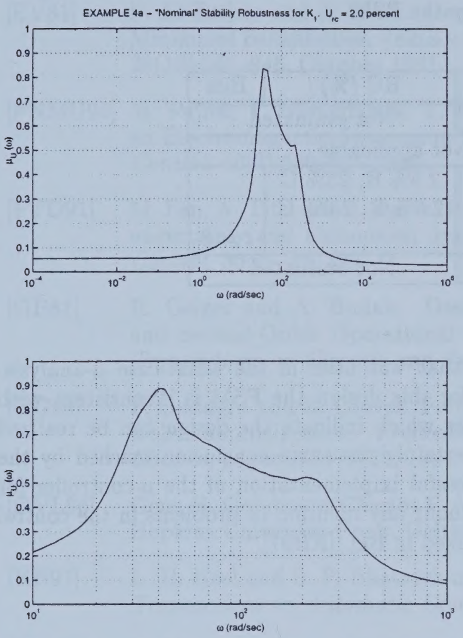


Figure 2: Closed-Loop controller robustness for Example # 4 showing nominal stability for $K_1(s)$.

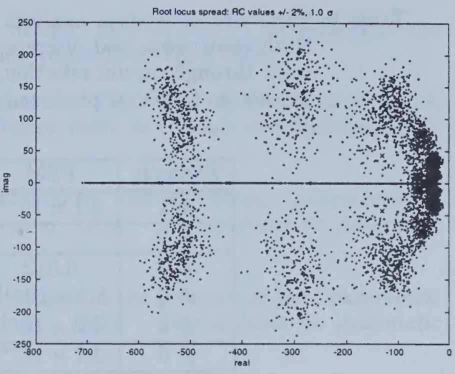


Figure 3: Closed-loop pole distribution ($1\sigma = 2\%$) for Example #4.

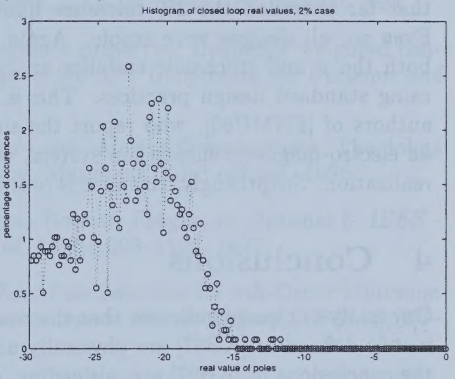


Figure 4: Histogram of closed-loop pole locations($1\sigma = 2\%$) for Example #4.

that this μ–controller can be built and is guaranteed to be stable using a total of 47 components (resistors and capacitors) that have parameter tolerances of ±2%. Figure (2) shows the μ upper bound for a coarse-grid and fine-grid frequency range. The peak value of the upper bound is approximately 0.9, which guarantees robust stability to the modeled variations (which in this case were independent ±2%, uniformly for all components). The Monte Carlo stability analysis for this particular control circuit gives an even more optimistic prediction of stability. For this analysis it was assumed that the 47 components were chosen randomly from a pool of resistors and capacitors with parameter values with a normal distribution, and a standard deviation $\sigma = 2\%$ of the nominal value. For the 1000 separate designs considered, the resulting closed-loop poles are plotted in Figure 3. As shown, all 1000 random designs were stable. In fact, the real component of the smallest eigenvalue never exceeded −11 rad/sec, or about one-half the nominal closed-loop value of −25 rad/sec. Indeed, the distribution of low frequency poles is quite tightly grouped about the nominal value (as shown in the histogram in Figure 4).

These results indicate that the design is robust to choice of components having standard deviations that are 2% of the nominal parameter values. Note that since a normal distribution was used each particular design of the 1000 had many component values

Table 1: Summary of modern analysis results for the six examples in Ref. [KB97]. Both continuous and discrete analyzes are presented. The results suggest that, through careful selection of the controller implementation, the designs are not as fragile as predicted by the PSM.

Example	PSM	RC (%)	Bits
# 1	2.1×10^{-7}	Not evaluated	
# 2	Not applicable		
# 3	0.012	2.5% R, 25% C	
# 4	1.4×10^{-15}	2.0% R, 2.0% C	
# 5	3.2×10^{-5}		$t = 8$
# 6	3.7×10^{-6}	Not evaluated	

that far exceeded the 2% tolerance limit that was used in the worst-case μ-analysis. Even so, all designs were stable. Again, for this design the PSM is inconsistent with both the μ and stochastic stability analyzes which indicate the design can be realized using standard design practices. This is certainly the same conclusion reached by the authors of [FNMU95], who report the sucessful implementation of the μ-controller for an electro-magnetic suspension system, without any mention of problems in the control realization. Surprisingly, this fact is overlooked in Ref. [KB97].

4 Conclusions

Our analysis clearly indicates that the controller uncertainty models and robust stability metric defined in [KB97] are physically unrealistic and typically pessimistic. As a result the conclusions in [KB97] are misleading, and are not consistent with control engineering practice and experience. With the high-level of interest in this topic that was demonstrated at the 1998 ACC session **FA09**, it is important that we correct the misconception that the optimizing control approaches such as LQG, $\mathcal{H}_\infty$, and μ result in controllers that are very difficult to implement. In this paper we take the first steps towards this goal by describing more realistic models of controller uncertainty, and presenting tools to perform the robustness analysis of the closed-loop systems with respect to these models. In the context of the examples that are often cited, our results indicate that the controllers could be successfully implemented using only modestly accurate analog componentry or typical floating point arithmetic. These results are in direct contrast with the degree of fragility predicted in Ref. [KB97].

References

[BCK95] S. Bhattacharyya, H. Chappellat, and L. Keel. *Robust Control: The Parametric Approach.* Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[DFT92] John C. Doyle, Bruce A. Francis, and Allen R. Tannenbaum. *Feedback Control Theory.* Macmillan Publishing Company, New York, New York, 1992.

[Dor98] P. Dorato. Non-fragile controller design: an overview. Proceedings of the *IEEE American Control Conference*, pages 2829–2831, 1998.

[EV81] F. El-Turkey and J. Vlach. Generation of Equivalent Active Networks with Minimized Sensitivities. *IEEE - Transactions on Circuits and Systems*, CAS-28(10):941–946, October 1981.

[FNMU95] M. Fujita, T. Namerikawa, F. Matsumura, and K. Uchida. μ-Synthesis of an Electromagnetic Suspension System. *IEEE - Transactions on Automatic Control*, 40(3):530–536, 1995.

[FTD91] M. Fan, A. Tits, and J. Doyle. Robustness in the presence of joint parametric uncertainty and unmodeled dynamics. *IEEE - Transactions on Automatic Control*, 36:25–38, 1991.

[GB81] R. Geiger and A. Budak. Design of Active Filters Independent of First- and Second-Order Operational Amplifier Time Constant Effects. *IEEE - Transactions on Circuits and Systems*, CAS-28(8):749–757, August 1981.

[GG94] P. Goddard and K. Glover. Performance preserving frequency weighted controller approximation. *Proceedings of the* 33^{th} *Conference on Decision and Control*, pages 2720–2725, 1994.

[GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, Maryland, third edition, 1996.

[KB97] L. H. Keel and S. P. Bhattacharyya. Robust, Fragile, or Optimal ? *IEEE - Transactions on Automatic Control*, 42(8):1098–1105, 1997.

[Mos70] G. Moschytz. Second-Order Pole-Zero Pair Selection for nth-Order Minimum Sensitivity Networks. *IEEE - Transactions on Circuit Theory*, CT-17(4):527–534, November 1970.

[Par71] S. Parker. Sensitivity: Old Questions, New Answers. *IEEE - Transactions on Circuit Theory*, CT-18(1):27–35, January 1971.

[RG71] A. Rosenblum and M. Ghausi. Multiparameter Sensitivity in Active RC Networks. *IEEE - Transactions on Circuit Theory*, CT-18(6):592–599, November 1971.

[RS93] L.R. Ray and Robert F. Stengel. A Monte Carlo Approach to the Analysis of Control System Robustness. *Automatica*, 29(1):229–236, January 1993.

[SE75] A. Sedra and J. Espinoza. Sensitivity and Frequency Limitations of Biquadratic Active Filters. *IEEE - Transactions on Circuits and Systems*, CAS-22(2):122–127, February 1975.

[SM71] M. Soderstrand and S. Mitra. Gain and Sensitivity Limitations of Active RC Filters. *IEEE - Transactions on Circuit Theory*, CT-18(6):600–609, November 1971.

[YP95] X. Yang and A. Packard. A low order controller design method. *Proceedings of the* 34^{th} *Conference on Decision and Control*, pages 3068–3073, 1995.

Friction Compensation in Mechanical Systems via Iterative Learning

Seong-Il Cho, Seung-Jean Kim, Young-Hoon Kim, and In-Joong Ha

School of Electrical Engineering, Seoul National University
San 56-1 Shinrim-Dong, Kwanak-Ku, Seoul, Korea 151-742
ijha@asri.snu.ac.kr

Abstract

This note describes a novel learning control scheme for tracking periodic trajectories in mechanical systems with friction. It is based on the fact that the solution of the closed-loop system tends to be periodic in steady state. When the closed-loop system reaches the steady state, the proposed learning control scheme updates the control input. By doing this iteratively, the proposed learning control scheme eventually can drive the tracking error to zero. Neither the information of the system mass nor the parametric model for friction is required for successful tracking. In particular, the proposed learning control scheme is computationally quite simple. Furthermore, its generality is well supported through rigorous convergence analysis.

1 Introduction

In motion control systems, friction is a primary source of disturbance, which can degrade control performance significantly. In this context, various friction compensation methods have been considered by many authors [2]. One of such friction compensation methods is the model-based friction compensation approach taken in [3]-[9]. It seems well suited for the case that an accurate parametric model for friction is available. In reality, however, it is often hard to obtain an accurate parametric model for friction. For this reason, a model-free design methodology was taken in [10], in which state-dependent parasitic effects such as friction were approximated by Gaussian neural networks. However, its main drawback is computational complexity.

On the other hand, it is well known that a learning approach is quite effective in the case that *a priori* knowledge of the system model is limited. In fact, learning control schemes have been popularly used in robotic manipulators, gas-metal arc welding process, and CNC machines [11]-[16]. Furthermore, the feasibility of a learning control scheme for friction compensation has been well demonstrated in [15] through experiments using a repetitive control scheme. The learning control schemes are usually computationally simple and hence can be implemented at cheap cost. Instead, iterative operations along with some memory are required.

In this note, we also take a learning approach to friction compensation. Our learning control scheme extends basically the work in [16] to the problem of periodic-trajectory tracking in mechanical systems with friction. In particular, our learning control scheme differs fundamentally from the repetitive control schemes [14], [15] and the conventional iterative learning schemes [11]-[13] in that its update scheme is based on the fact that the solution of the closed-loop system tends to be periodic in steady state. Compared to other friction compensation methods, ours has the following merits. First, the proposed learning control scheme is computationally quite simple. Second, it is shown through a mathematically rigorous analysis on discontinuous nonlinear systems that the tracking error eventually converges to zero in a few number of iterations but neither the accurate information of the system mass nor the parametric model for friction is required for the convergence of the tracking error.

In our development, the usual Euclidean norm of a vector $x \in \mathcal{R}^n$ is denoted by $\|x\|$. On the other hand, a function $f : [0, \infty) \to \mathcal{R}$ is called T-periodic if $f(t+T) = f(t)$, $\forall t \geq 0$, for a constant $T > 0$. Finally, the inner product and the norm on $L_2[0,\ T]$ are denoted by $\langle f|g\rangle \triangleq \int_0^T f(t)g(t)dt$ and $\|f\|_T \triangleq \langle f|f\rangle^{1/2}$, respectively.

2 Preliminary Result

The dynamics of the mechanical system with friction considered in this note are described by the following 2nd-order differential equation.

$$m\ddot{x}(t) \;+\; F(\dot{x}(t), u(t)) \;=\; u(t), \tag{1}$$

where m is a mass, $x(t) \in \mathcal{R}$ is the position, F is the friction force, and $u(t) \in \mathcal{R}$ is the control input. Here, m and F are unknown. However, the structure of the friction force is assumed to be *a priori* known as follows [7].

$$F(\dot{x}, u) = F_s(\dot{x}) + F_0(\dot{x}, u). \tag{2}$$

Here, the function F_s represents the friction force at non-zero velocity and is assumed to be modeled as follows.

$$F_s(\dot{x}) \;= B\dot{x} \;+\; F_n(\dot{x}), \tag{3}$$

where B is the viscous friction coefficient and $F_n(\dot{x})$ is the nonlinear function of $\dot{x}$ which is continuous except at zero velocity. The typical form of the function F_s is shown in Fig.1.

On the other hand, the function F_0 represents the friction force at zero velocity and its mathematical model is given by

$$F_0(\dot{x}, u) \;\triangleq\; \begin{cases} 0 & \text{if } \dot{x} \neq 0 \\ \lambda(u) & \text{if } \dot{x} = 0 \end{cases}, \tag{4}$$

where

$$\lambda(u) \;\triangleq\; \begin{cases} F_s(0+) & \text{if} \quad u \quad \geq F_s(0+) > \;0 \\ u & \text{if } F_s(0-) < \quad u \quad < F_s(0+) \\ F_s(0-) & \text{if} \quad u \quad \leq F_s(0-) < \;0 \end{cases}. \tag{5}$$

For mathematical tractability, we assume that the function F_s in (3) has the following properties, which have been accepted commonly in the prior literature. See, for example, [2].

$$\frac{F_s(\dot{x}_1) - F_s(\dot{x}_2)}{\dot{x}_1 - \dot{x}_2} \geq S_b, \qquad \text{if} \qquad \dot{x}_1 \neq \dot{x}_2, \tag{6}$$

$$|F_n(\dot{x})| \;\leq\; F_b,\ \forall \dot{x} \in \mathcal{R} \quad \text{and} \quad F_n(0) = 0, \tag{7}$$

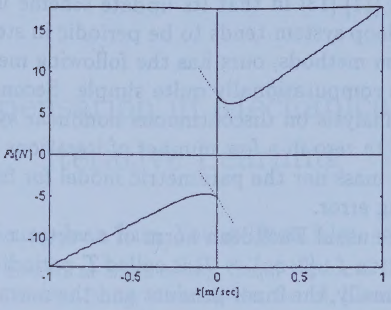


Figure 1: Friction Curve

where S_b and F_b are constants. In fact, many empirical models including the Gaussian model and the Lorentzian model [2] can be easily shown to satisfy the above assumptions.

Our tracking controller consists of the well-known proportional-plus-derivative (PD) controller and a feedforward input, say u^f, for error correction as follows.

$$u(t) \;=\; K_p(x^d(t) - x(t)) \;+\; K_d(\dot{x}^d(t) - \dot{x}(t)) \;+\; u^f(t), \tag{8}$$

where K_p, K_d are positive constants. Here, the function $x^d : [0, \infty) \to \mathcal{R}$ represents the desired trajectory which is assumed to be T-periodic and twice continuously differentiable. On the other hand, the function $u^f : [0, \infty) \to \mathcal{R}$ is updated iteratively, as described later. The following Theorem 1 shows that as $t \to \infty$, any solution of the closed-loop system given by (1) and (8) approaches a T-periodic solution.

Theorem 1 (Existence of Steady-State Oscillation) *Suppose that u^f in (8) is continuous and T-periodic. If the controller gains K_p, K_d in (8) are chosen so as to satisfy*

$$K_p > 0,\; K_d + S_b > 0, \tag{9}$$

then any solution, say $z(t) \triangleq [x(t)\; \dot{x}(t)]^T \in \mathcal{R}^2$, of the closed-loop system given by (1) and (8) converges to an absolutely continuous T-periodic solution, say $\bar{z}(t) \triangleq [\bar{x}(t)\; \dot{\bar{x}}(t)]^T \in \mathcal{R}^2$ such that

$$\lim_{t\to\infty} \|z(t) - \bar{z}(t)\| = 0. \tag{10}$$

Note that by our definition of T-periodicity, the T-periodic solution $\bar{z}$ may be identically constant. As a matter of fact, a mechanical system with friction may continue to stay stuck in case of small control input. In what follows, the T-periodic solution $\bar{z}$ will be called *steady-state oscillation.* According to Theorem 1, the steady-state oscillation always occurs if the controller gains are chosen sufficiently large. Theorem 1 is very important since it provides the theoretical foundation for our learning approach. Nonetheless, the proof of Theorem 1 is omitted here because of limited space but will be presented elsewhere.

3 Main Result

To describe precisely the update scheme for u^f in (8), we need some notational convention. The subscript i is used to denote the iteration number, while u_i^f represents the *ith* update of the function u^f. Correspondingly, the subscript i is attached to all variables that are affected by letting $u^f = u_i^f$ in (8). The n_i are nonnegative integers. Note that we need to define the function u^f only for one period $[0,\ T]$ since it is T-periodic. Finally, the tracking error is denoted by $e(t) \triangleq x^d(t) - x(t)$. Now, we are ready to describe the iterative update scheme for u^f.

Step 1) Set $i = 0$. Let $u^f(\bar{t}) = u_0^f(\bar{t}) \triangleq 0,\ \forall \bar{t} \in [0,\ T]$.
Step 2) Wait until the closed-loop system given by (1) and (8) exhibits the steady-state oscillation. Then, store the time history of the velocity error $\dot{e}_i$ for one period, say, $[n_i T, (n_i + 1)T]$.
Step 3) Let $\bar{e}_i(\bar{t}) \triangleq e_i(\bar{t} + n_i T)$ and $\dot{\bar{e}}_i(\bar{t}) \triangleq \dot{e}_i(\bar{t} + n_i T)$ for all $\bar{t} \in [0,\ T]$. Determine u_{i+1}^f using

$$u_{i+1}^f(\bar{t}) \triangleq u_i^f(\bar{t}) + \gamma \dot{\bar{e}}_i(\bar{t}) + \frac{K_p}{T} \int_0^T \bar{e}_i(\tau) d\tau, \quad \forall \bar{t} \in [0,\ T], \tag{11}$$

where γ is a constant satisfying

$$0 < \gamma < 2(S_b + K_d). \tag{12}$$

Step 4) At $t = (n_i + 1)T$, let $u^f = u_{i+1}^f$. Then, increase i by 1 and jump to *Step 2.*

The timing diagram of the proposed update scheme is depicted in Fig.2. Note that the function u^f is not updated continuously with respect to time. Instead, it is switched once at each iteration from the present waveform to a new waveform after the closed-loop system with $u^f = u_i^f$ reaches the steady state.

As can be seen from (11), there are two error correction terms in the update of u^f. The one is a T-periodic signal proportional to the derivative of the steady-state tracking error. As can be seen from the proof of the following Theorem 2, it plays a key role in driving the tracking error to a constant (which may not be zero) as the iteration number increases. The other is a constant proportional to the time-average value of the steady-state tracking error. As can be seen below, it is to assure that the tracking error goes necessarily to zero as the iteration number increases.

Now, we are ready to state our main result.

Theorem 2 (Convergence of Tracking Error) *Suppose that the feedforward input u^f in (8) is updated according to* Step 1) - Step 4). *Suppose further that the set*

$$\mathcal{N} \triangleq \{t \in [0,\ T] \mid \dot{x}^d(t) = 0\} \quad \text{is of measure zero.} \tag{13}$$

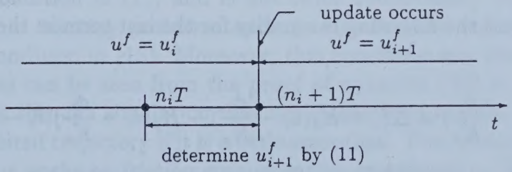


Figure 2: Timing diagram of the proposed update scheme.

Then,

$$\lim_{i\to\infty} \bar{e}_i(t) = 0, \quad \forall t \in [0,\ T]. \tag{14}$$

Proof) First, we prove that

$$\lim_{i\to\infty} \|\dot{\bar{e}}_i\|_T = 0. \tag{15}$$

Define a T-periodic function u^d by

$$u^d(t) \triangleq m\ddot{x}^d(t) + F_s(\dot{x}^d(t)). \tag{16}$$

Let $\bar{u}^d$ and $\bar{u}^f$ be the time-average components of u^d and u^f, respectively. Let $\tilde{u}^d \triangleq u^d - \bar{u}^d$ and $\tilde{u}^f \triangleq u^f - \bar{u}^f$. Then, we can rewrite (11) as follows.

$$\bar{u}^f_{i+1} = \bar{u}^f_i + \frac{K_p}{T}\int_0^T \bar{e}_i(\tau)d\tau, \quad \tilde{u}^f_{i+1}(t) = \tilde{u}^f_i(t) + \gamma\dot{\bar{e}}_i(t). \tag{17}$$

Define new variables ϕ and ξ by

$$\phi_i \ \triangleq\ \bar{u}^d - \bar{u}^f_i, \quad \xi_i(t) \ \triangleq\ \tilde{u}^d(t) - \tilde{u}^f_i(t). \tag{18}$$

By (17) and (18), we can obtain that

$$\xi_{i+1}(t) = \xi_i(t) - \gamma\dot{\bar{e}}_i(t), \forall t \in [0,T], \qquad \phi_{i+1} = \phi_i - K_p\bar{E}_i, \tag{19}$$

where $\bar{E}_i \triangleq \frac{1}{T}\int_0^T \bar{e}_i(\tau)d\tau$. Hence, we have

$$\|\xi_{i+1}\|_T^2 = \|\xi_i\|_T^2 + \gamma^2\|\dot{\bar{e}}_i\|_T^2 - 2\gamma\langle\xi_i|\dot{\bar{e}}_i\rangle. \tag{20}$$

On the other hand, we can obtain the following dynamic equation for the tracking error in steady state from (1), (8), (16), and (18).

$$m\ddot{\bar{e}}_i + F_s(\dot{x}^d) - F(\dot{x}^d - \dot{\bar{e}}_i,\ K_p\bar{e}_i + K_d\dot{\bar{e}}_i + u^d - \phi_i - \xi_i) + K_d\dot{\bar{e}}_i + K_p\bar{e}_i = \phi_i + \xi_i. \tag{21}$$

Since ϕ_i is a constant, this can be written as

$$m(\ddot{\bar{e}}_i - \frac{\ddot{\phi}_i}{K_p}) + \Delta F_i + K_d(\dot{\bar{e}}_i - \frac{\dot{\phi}_i}{K_p}) + K_p(\bar{e}_i - \frac{\phi_i}{K_p}) = \xi_i, \tag{22}$$

where $\Delta F_i \triangleq F_s(\dot{x}^d) - F(\dot{x}^d - (\dot{\bar{e}}_i - \frac{\dot{\phi}_i}{K_p}),\ K_p(\bar{e}_i - \frac{\phi_i}{K_p}) + K_d(\dot{\bar{e}}_i - \frac{\dot{\phi}_i}{K_p}) + u^d - \xi_i)$. Furthermore, we can easily show using (6) and Theorem 1 that

$$\langle\dot{\bar{e}}_i|\bar{e}_i\rangle \ = \ \langle\dot{\bar{e}}_i|\ddot{\bar{e}}_i\rangle \ = \ 0. \tag{23}$$

$$(\dot{x}_1 - \dot{x}_2)\left[F_s(\dot{x}_1) - F(\dot{x}_2, u) + K_d\dot{x}_1 - K_d\dot{x}_2\right] \ \geq \ (S_b + K_d)|\dot{x}_1 - \dot{x}_2|^2, \ \forall\dot{x}_1, \dot{x}_2, u \in \mathcal{R}. \tag{24}$$

By (22)-(24), we can find the following inequality for the last term in the Right-Hand Side (RHS) of (20).

$$\begin{aligned}\langle\xi_i|\dot{\bar{e}}_i\rangle \ &= \ \langle m(\ddot{\bar{e}}_i - \frac{\ddot{\phi}_i}{K_p}) + \Delta F_i + K_d(\dot{\bar{e}}_i - \frac{\dot{\phi}_i}{K_p}) + K_p(\bar{e}_i - \frac{\phi_i}{K_p})|\dot{\bar{e}}_i - \frac{\dot{\phi}_i}{K_p}\rangle \\ &= \ \langle\Delta F_i + K_d(\dot{\bar{e}}_i - \frac{\dot{\phi}_i}{K_p})|\dot{\bar{e}}_i - \frac{\dot{\phi}_i}{K_p}\rangle \quad \geq \ (S_b + K_d)\|\dot{\bar{e}}_i\|_T^2.\end{aligned} \tag{25}$$

Now, the following inequality is the direct consequence of (12), (20), and (25).

$$\|\xi_{i+1}\|_T^2 - \|\xi_i\|_T^2 \leq -\gamma\{2(S_b + K_d) - \gamma\}\|\dot{\bar{e}}_i\|_T^2 \leq 0, \quad \forall i = 0, 1, 2, \cdots. \tag{26}$$

By induction using this fact, we can show that

$$\gamma\{2(S_b + K_d) - \gamma\} \sum_{i=0}^{\infty} \|\dot{\bar{e}}_i\|_T^2 \leq \|\xi_0\|_T^2 < \infty. \tag{27}$$

This clearly implies (15).

It is shown in the Appendix that if (15) holds, then

$$\lim_{i\to\infty} \frac{1}{T} \int_0^T \left(\bar{F}_i(\tau) - \bar{F}_{i-1}(\tau)\right) d\tau = 0, \tag{28}$$

where $\bar{F}_i(\tau) \triangleq F(\dot{\bar{x}}_i(\tau),\ K_p\bar{e}_i(\tau) + K_d\dot{\bar{e}}_i(\tau) + u^d(\tau) - \phi_i - \xi_i(\tau))$.

Now, we prove the desired property in (14) using this fact. For this aim, we first integrate both sides of (21) over $[0,\ T]$. As the result, we have

$$\frac{1}{T}\int_0^T (F_s(\dot{x}^d(\tau)) - \bar{F}_i(\tau))d\tau + K_p\bar{E}_i \;=\; \frac{1}{T}\int_0^T (\phi_i + \xi_i(\tau))d\tau \;=\; \phi_i. \tag{29}$$

By this along with (19),

$$\phi_{i+1} = \frac{1}{T}\int_0^T (F_s(\dot{x}^d(\tau)) - \bar{F}_i(\tau))d\tau. \tag{30}$$

Therefore, we have

$$K_p\bar{E}_i = \frac{1}{T}\int_0^T (\bar{F}_i(\tau) - \bar{F}_{i-1}(\tau))d\tau. \tag{31}$$

By (28), this implies that

$$\lim_{i\to\infty} \bar{E}_i = 0. \tag{32}$$

For two arbitrary constants $t, t_0 \in [0,\ T]$, we define the function $\hat{E}_i : \mathcal{R}^2 \to \mathcal{R}$ by $\hat{E}_i(t, t_0) \triangleq \int_{t_0}^t \dot{\bar{e}}_i(\tau)d\tau$. By the definitions of $\bar{E}_i$ and $\hat{E}_i$, we then have

$$\bar{E}_i = \frac{1}{T}\int_0^T \hat{E}_i(\tau, t_0)d\tau + \bar{e}_i(t_0). \tag{33}$$

By (15) along with the Schwarz Inequality, we also have

$$\lim_{i\to\infty} \hat{E}_i(t, t_0) = 0, \quad \forall t, t_0 \in [0\ \ T]. \tag{34}$$

Now, our claim in (14) is the immediate consequence of (32), (33), and (34).

The condition in (13) requires that the desired trajectory should not be identically constant for a time interval. For instance, a square-waveform desired trajectory does not satisfy the condition in (13) and is not twice continuously differentiable. Nonetheless, many dynamic tracking problems such as circular end milling in CNC machines can satisfy the condition in (13). Moreover, this condition was introduced for mathematical tractability as can be seen from the proof of equation (28) in the Appendix. In fact, it can be shown through simulation that the closed-loop system can track well the square-waveform desired trajectory if it is a little smoothed. Furthermore, it should be noted that even the prior works on friction compensation mentioned in Section I can be supported theoretically well under the conditions similar to that in (13).

4 Simulation Results

For our simulation purpose, the function F_s in (2) was assumed to be expressed by

$$F_s(\dot{x}) \triangleq F_s^+(\dot{x})\nu(\dot{x}) + F_s^-(\dot{x})[1 - \nu(\dot{x})], \tag{35}$$

where $\nu(\dot{x}) \triangleq \begin{cases} 1 & \text{if } \dot{x} > 0 \\ 0 & \text{if } \dot{x} < 0 \end{cases}$. $F_s^+(\dot{x}) \triangleq B^+\dot{x} + F_c^+ + (F_z^+ - F_c^+)e^{-|\dot{x}/\dot{x}_s^+|}$, and $F_s^-(\dot{x}) \triangleq B^-\dot{x} + F_c^- + (F_z^- - F_c^-)e^{-|\dot{x}/\dot{x}_s^-|}$, where $B^+ = 13.7[N\text{sec}/m]$, $B^- = 10.5[N\text{sec}/m]$, $F_c^+ = 4.2[N]$, $F_c^- = -3.0[N]$, $F_z^+ = 7.3[N]$, $F_z^- = -5.0[N]$, $\dot{x}_s^+ = 0.07[m/\text{sec}]$, and $\dot{x}_s^- = 0.09[m/\text{sec}]$. The friction curve corresponding to the above data can be plotted graphically as shown in Fig.1. Observe from Fig.1 that the asymmetry of friction as well as the Stribeck effect has been considered fully. On the other hand, we chose x^d, m, K_p, K_d, and γ as follows.

$$x^d(t) = sin(\pi t)[m], \quad m = 1.0[kg], \quad K_p = 120, \quad K_d = 80, \quad \gamma = 70. \tag{36}$$

To verify Theorem 1 first, we have investigated the dynamic behavior of the closed-loop system given by (1) and (8). The simulation results for three different initial conditions given in Fig.3 clearly show that the closed-loop system tends to oscillate in steady state. To verify Theorem 2 next, u^f in (8) was updated according to *Step 1) - Step 4)*. To confirm that the closed-loop system reaches the steady state at each iteration, the time interval when the update occurs was chosen sufficiently large to be 8 [sec]. The simulation results are shown in Fig.4, where the encircled number ⓘ stands for the *ith* iteration. Fig.4 shows that the proposed learning control scheme drives $\bar{e}$ and $\dot{\bar{e}}$ to almost zero in six iterations.

5 Conclusion

To our best knowledge, this is the first result to present a mathematically rigorous analysis on iterative learning control of discontinuous nonlinear systems. Nonetheless, the proposed learning control scheme has some drawbacks to resolve in further research. First, it is necessary to perform different trials for each of new desired trajectories. Second, the convergence rate may be slow since each update occurs only after the closed-loop system reaches the steady state. Third, the proposed method cannot handle output regulation problems or set-point tracking problems effectively. Finally, the proposed update scheme requires some memory to store the feedforward input for one period.

Nonetheless, the proposed scheme can be implemented easily at low cost. Furthermore, neither the information of the system mass nor the parametric model for friction is required for successful tracking. As a matter of fact, we have shown that the generality and feasibility of the proposed scheme are well supported through rigorous convergence analysis and simulation results. Hence, it should find a wide range of applications in industries.

Appendix

Proof of (28)

Define the subsets $\mathcal{O}, \mathcal{A}_i, \mathcal{B}_i, \mathcal{C}_i$ of the interval $[0,T]$ as follows. $\mathcal{O} \triangleq \{t \in [0,T] | \lim_{i\to\infty} \dot{x}_i(t) \neq \dot{x}^d(t)\}$, $\mathcal{A}_i \triangleq \{t \in [0,T] | \dot{x}_i(t)\dot{x}_{i-1}(t) = 0\} - \mathcal{O}$, $\mathcal{B}_i \triangleq \{t \in [0,T] | \dot{x}_i(t)\dot{x}_{i-1}(t) < 0\} - \mathcal{O}$, and $\mathcal{C}_i \triangleq \{t \in [0,T] | \dot{x}_i(t)\dot{x}_{i-1}(t) > 0\} - \mathcal{O}$. Then, $[0,\ T] = \mathcal{O} \bigcup \mathcal{A}_i \bigcup \mathcal{B}_i \bigcup \mathcal{C}_i, \quad \forall i = 1, 2, 3, \cdots$. Therefore, we can write

$$\begin{aligned}\int_0^T (\bar{F}_i - \bar{F}_{i-1})d\tau = \int_{\mathcal{O}} (\bar{F}_i - \bar{F}_{i-1})d\tau &+ \int_{\mathcal{A}_i} (\bar{F}_i - \bar{F}_{i-1})d\tau \\ &+ \int_{\mathcal{B}_i} (\bar{F}_i - \bar{F}_{i-1})d\tau + \int_{\mathcal{C}_i} (\bar{F}_i - \bar{F}_{i-1})d\tau. \quad \text{(A.1)}\end{aligned}$$

Firstly, we evaluate the first two terms in the RHS of (A.1). In what follows, $\mu(\mathcal{S})$ denotes the Lebesgue measure of a set $\mathcal{S}$ [1]. Note from (13) and (15) that $\mu(\mathcal{O}) = 0$ and $\lim_{i \to \infty} \mu(\mathcal{A}_i) = 0$. Therefore, we have

$$\lim_{i \to \infty} \left\{ \int_{\mathcal{O}} (\bar{F}_i(\tau) - \bar{F}_{i-1}(\tau))d\tau + \int_{\mathcal{A}_i} (\bar{F}_i(\tau) - \bar{F}_{i-1}(\tau))d\tau \right\} = 0. \qquad \text{(A.2)}$$

Secondly, we evaluate the third term in the RHS of (A.1). Choose an arbitrary constant $\epsilon > 0$. Then, define the sets $\mathcal{P}_\epsilon$, $\mathcal{N}_\epsilon$, and $\mathcal{S}_\epsilon$ as follows. $\mathcal{P}_\epsilon \triangleq \{t \in [0,\ T] \mid \dot{x}^d(t) \geq \epsilon\}$, $\mathcal{N}_\epsilon \triangleq \{t \in [0,\ T] \mid \dot{x}^d(t) \leq -\epsilon\}$, and $\mathcal{S}_\epsilon \triangleq \{t \in [0,\ T] \mid |\dot{x}^d(t)| < \epsilon\}$. By Proposition 24 in [1] along with (15), there is a subset $\mathcal{E}$ of $[0,\ T]$ and an integer N such that

$$\mu(\mathcal{E}) < \epsilon, \quad |\dot{x}_{i-1} - \dot{x}^d| < \epsilon,\ \forall t \in [0,\ T] - \mathcal{E},\ \forall i \geq N. \qquad \text{(A.3)}$$

Furthermore, it follows from (A.3) and the definition of $\mathcal{B}_i$ that if $i \geq N$, $t \in \mathcal{B}_i$ implies that $t \notin (\mathcal{P}_\epsilon - \mathcal{E}) \bigcup (\mathcal{N}_\epsilon - \mathcal{E})$ and hence that $t \in \mathcal{S}_\epsilon \cup \mathcal{E}$. Therefore, we conclude that

$$\mathcal{B}_i \subset \mathcal{S}_\epsilon \cup \mathcal{E},\ \forall i \geq N. \qquad \text{(A.4)}$$

On the other hand, we see from (13) that $\lim_{\epsilon \to 0} \mu(\mathcal{S}_\epsilon) = 0$. By (A.3) and (A.4), this implies that $\lim_{i \to \infty} \mu(\mathcal{B}_i) = 0$. Therefore, we have

$$\lim_{i \to \infty} \int_{\mathcal{B}_i} (\bar{F}_i(\tau) - \bar{F}_{i-1}(\tau))d\tau = 0. \qquad \text{(A.5)}$$

Finally, we evaluate the last term in the RHS of (A.1). Note that the above function $\bar{F}_i - \bar{F}_{i-1}$ is continuous at each $t \in \mathcal{C}_i$. Therefore, for any $\epsilon > 0$, there exists an integer N such that $|\bar{F}_i(t) - \bar{F}_{i-1}(t)| < \epsilon,\ \forall t \in \mathcal{C}_i, \quad \forall i \geq N$. This implies that

$$\lim_{i \to \infty} \int_{\mathcal{C}_i} (\bar{F}_i(\tau) - \bar{F}_{i-1}(\tau))d\tau = 0. \qquad \text{(A.6)}$$

Now, (28) is the immediate consequence of (A.2), (A.5), and (A.6).

References

[1] H. L. Royden, *Real Analysis, Macmillan Publishing Company*, third edition, 1989.

[2] B. Armstrong-Hélouvry, P. Dupont, and C. Canudas de Wit, "A Survey of Models, Analysis Tools and Compensation Methods for the Control of Machines with Friction", *Automatica*, vol. 30, no. 7, pp. 1083-1138, 1994.

[3] J. W. Gilbart and G. C. Winston, "Adaptive Compensation for an Optical Tracking Telescope" *Automatica*, vol. 10, no. 2, pp. 125-131, 1974.

[4] C. D. Walrath, "Adaptive Bearing Friction Compensation Based on Recent Knowledge of Dynamic Friction," *Automatica*, vol. 20, no. 6, pp. 717-727, 1984.

[5] C. Canudas de Wit, K. J. Åström and K. Braun, "Adaptive Friction Compensation in DC-Motor Drives," *IEEE J. Robotics and Automation*, vol. 3, no. 6, pp. 681-685, Dec, 1987.

[6] B. Friedland and Y. J. Park, "On Adaptive Friction Compensation", *IEEE Trans. Automat. Contr.*, vol. 37, no. 10, pp. 1609-1612, Oct, 1992.

[7] S. W. Lee and J. H. Kim, "Robust Adaptive Stick-Slip Friction Compensation", *IEEE Industrial Electronics*, vol. 42, no. 5, pp. 474-479, Oct, 1995.

[8] J. Amin, B. Friedland, and A. Harnoy, "Implementation of a Friction Estimation and Compensation Technique", *IEEE Control Systems Magazine*, vol. 17, no. 4, pp. 71-76, Aug, 1997.

[9] C. G. Baril and P. Gutman, "Performance Enhancing Adaptive Friction Compensation for Uncertain Systems", *IEEE Control Systems Technology*, vol. 5, no. 5, pp. 466-479, Sep, 1997.

[10] H. Du and S. S. Nair, "Low Velocity Friction Compensation", *IEEE Control Systems Magazine*, vol. 18, no. 2, pp. 61-69, Apr, 1998.

[11] S. Arimoto, S. Kawamura, F. Miyazaki and S. Tamaki, "Learning Control Theory for Dynamical Systems", *Proc. of IEEE 24th CDC*, vol. 3, pp. 1375-1380, 1985.

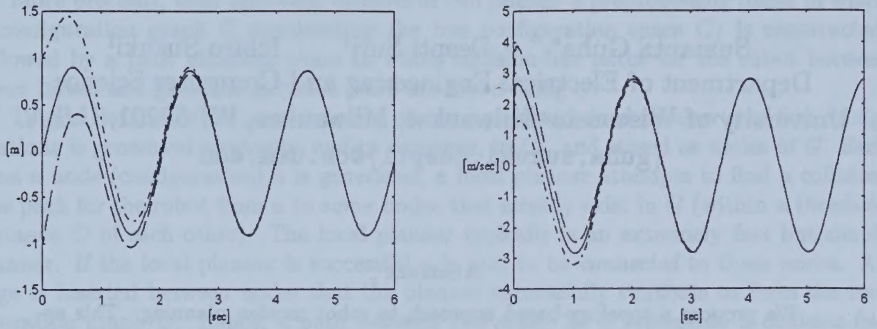
[12] P. Bondi, G. Casalino, and L. Gambardella, "On the Iterative Learning Control Theory for Robot Manipulators", *IEEE J. Robotics and Automation*, vol. 4, pp. 14-22, 1988.

[13] K. L. Moore and A. Mathews, "Iterative Learning Control for Systems with Non-Uniform Trial Length with Applications to Gas-Metal Arc Welding", *Proceeding of the 2nd Asian Control Conference.*, vol. 2, pp. 35-38, 1997.

[14] S. Hara, Y. Yamamoto, T. Omata, and M. Makano, "Repetitive Control Systems: A New Type Servo System for Periodic Exogeneous Signals," *IEEE Trans. Automat. Contr.*, vol. 33, no. 7, pp. 659-667, 1988.

[15] E. D. Tung, G. Anwar, and M. Tomizuka, "Low Velocity Friction Compensation and Feedforward Solution Based on Repetitive Control", *ASME J. Dynamic Syst., Meas. and Contr.*, vol. 115, pp. 279-284, 1993.

[16] I. J. Ha, S. H. Han, and Y. H. Kim, "Iterative Identification of State-Dependent Disturbance Torque for High Precision Velocity Control of Servo motors", *IEEE Trans. Automat. Contr.* pp. 724-729, May, 1998.

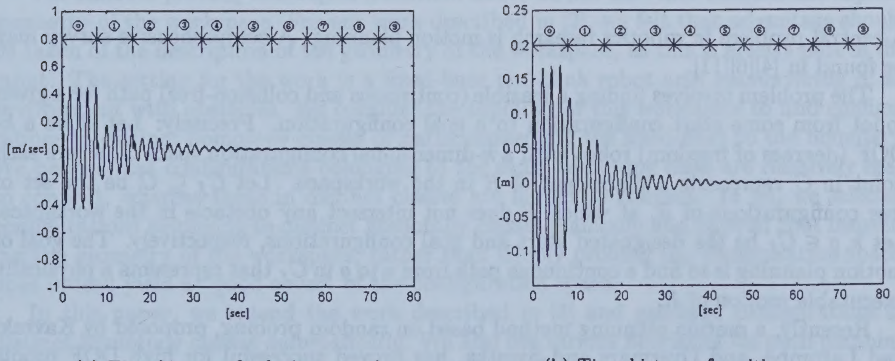


(a) Steady-state response of position x

(b) Steady-state response of velocity $\dot{x}$

Solid line: $(x(0), \dot{x}(0))=(0,0)$, Dashed line: $(x(0), \dot{x}(0))=(-1,-1)$, Dot-Dashed line: $(x(0), \dot{x}(0))=(1,1)$

Figure 3: Steady-state characteristics of the closed-loop system($u^f \equiv 0$)



(a) Time history of velocity error $\dot{e}$

(b) Time history of position error e

Figure 4: Effect of the iterative update scheme

Topology-Guided Motion Planning

Sumanta Guha* Deepti Suri† Ichiro Suzuki‡
Department of Electrical Engineering and Computer Science
University of Wisconsin–Milwaukee, Milwaukee, WI 53201, U.S.A.
`{guha,suzuki,deepti}@cs.uwm.edu`

Abstract

We propose a topology-based approach to robot motion planning. This approach complements a recently proposed and very successful method using randomization. We focus on implementing our ideas on a 7-DOF robot manipulator. Our scheme is to make explicit use of the geometry of the workspace in which the manipulator operates to identify a set of configurations that forms a "good cover" of the entire free configuration space. Our method automatically generates this cover. The configurations in this cover are from distinct homotopy classes of paths in the workspace. We run simulation experiments that indicate that a topology-based approach, informed by the geometry of the workspace, performs as well as the approach based on "blind" random probing of the workspace. However, we feel that exploiting the topology of the workspace has added advantages including: simplicity, reduction in the size of the configuration graph, and adaptability to a dynamic environment.

1 Introduction

A central problem in robotics research is motion planning. A comprehensive survey may be found in [4][6][11].

The problem involves finding a feasible (continuous and collision-free) path for a given robot from some start configuration to a goal configuration. Precisely: Let R be a k-DOF (degrees of freedom) robot with a k-dimensional configuration space C, where each point in C represents a placement of R in the workspace. Let $C_f \subseteq C$ be the set of *free* configurations of R, at which R does not intersect any obstacle in the workspace. Let $s, g \in C_f$ be the designated start and goal configurations, respectively. The goal of motion planning is to find a continuous path from s to g in C_f that represents a physically executable motion of R.

Recently, a motion planning method based on random probing, proposed by Kavraki and Latombe, and Overmars and Švestka, has proved successful for high DOF robots [5][8]. Their basic idea is to randomly generate points in the free configuration space C_f

---

[*] Supported in part by the National Science Foundation under grant CCR-9710771.

[†] Communicating Author

[‡] Supported in part by the National Science Foundation under grant IRI-9307506, the Office of Naval Research under grant N00014-94-1-0284, and Hitachi Ltd.

and attempt to connect these points by paths in C_f into a network, called the *configuration graph*, using a simple and efficient local planner. The goal is that the points of the configuration graph be a good "cover" of C_f, in the sense that an arbitrary point in C_f can be connected to some point in this cover by the local planner.

More precisely, their approach consists of two phases: a preprocessing phase in which a configuration graph G representing the free configuration space C_f is constructed, followed by a path planning phase in which collision-free paths for the robot between given initial and goal configuration pairs are found using G.

To give details, in the preprocessing phase, a predetermined number n of free configurations is generated randomly, and in sequence, in C_f, and stored as nodes of G. Each time a node (configuration) a is generated, a local planner attempts to find a collision-free path for the robot from a to *some* nodes that already exist in G (within a threshold distance D of each other). The local planner typically is an extremely fast but simple planner. If the local planner is successful a is said to be *connected* to these nodes. An edge is inserted between nodes that the planner successfully connects to form the configuration graph G. Hence, a path between two nodes in G represents a collision-free motion between the corresponding two configurations. G may, of course, consist of more than one component.

Subsequently, in the path planning phase one first attempts to connect, using the same local planner, the start configuration s and the goal configuration g to nodes s' and g' of G, respectively. If this is successful one then tries to find a path joining s' and g' in G. Of course, such a path exists if and only if s' and g' belong to the same component of G.

In [1][2] the authors investigate *safe* motion planning. In [2], we adapt the random probing technique to approximate medial axes and plan safe motion. It should be noted that, though the computation of medial axes is a classical problem e.g., in image processing [9], the application of random probing techniques is novel. Our investigations into safe motion planning using random probing to approximate the medial axis led us to consider topological methods that are subject of this paper.

The random probing techniques described above do not take into account the explicit geometry of the workspace. For our work described in [3], we felt that advantage should be taken of the description of the geometry of the workspace, as this is always part of the input. The setting for the work is a fixed-base multilink robot arm maneuvering in an environment with obstacles. We consider the topology of this workspace to find configurations in distinct homotopy classes of paths connecting the fixed-base to the boundary. We believe that configurations belonging to the same homotopy class are relatively easy to connect, whereas those in distinct classes are harder to connect. In [3], we selected configurations belonging to distinct homotopy classes manually and showed that inspecting the geometry of the workspace, rather than simply probing the configuration space, does indeed yield a "good cover" of the configuration space.

In this paper, we extend the work described in [3] and generate configurations in distinct homotopy classes *automatically*. We also use various heuristics for path planning to connect between configurations that belong to different homotopy classes using the local planner.

Our experiments indicate that our approach is indeed as good as the random probing of the workspace with the added advantages of simplicity, reduction in size of the configuration graph, and stability under small perturbations of the workspace.

2 Applications of Topology to Motion Planning

In the following section we first discuss some mathematical prerequisites. The material is from [3]. Further sources include [7][10].

2.1 Mathematical Prerequisites

2.1.1 Topology

A topology on a set X is a collection τ of subsets of X such that

1. $\emptyset, X \in \tau$,
2. if $\{U_i\} \subseteq \tau$ then $\bigcup U_i \in \tau$, and
3. if $\{U_i\} \subseteq \tau$ is finite then $\bigcap U_i \in \tau$.

A set X is with a topology τ is called a *topological space.* A set in τ is an *open* set, while a set whose complement is open is called *closed.*

A open *cover* of X is a collection of open sets whose union is X.

X is *connected* if the only sets both open and closed in X are the empty set and X itself. A *component* of X is a maximal connected subset.

A function $f : X \to Y$ from one topological space to another is *continuous* if the preimage of every open set in Y is open in X. A *map* is a continuous function. A *homeomorphism* is a bijective map whose inverse is also continuous. X and Y are *homeomorphic* or *topologically equivalent*, denoted $X \sim Y$, if there is a homeomorphism between them. A map $f : X \to Y$ is an *embedding* if it maps X homeomorphically onto its image $f(X)$ in Y.

A *homotopy* between two maps $f, g : X \to Y$ is a map

$$F : X \times [0,1] \to Y$$

for which the *initial map* $F(x,0) = f(x)$ and the *end map* $F(x,1) = g(x)$ for all $x \in X$. Two maps f and g are *homotopic* if such an F exists. Two topological spaces, X and Y, are *homotopy equivalent*, denoted $X \approx Y$, if there are maps $f : X \to Y$ and $g : Y \to X$ with $g \circ f$ homotopic to the identity map 1_X and $f \circ g$ homotopic to 1_Y.

Let $I = [0,1]$ be the unit interval. Then, a *path* in T from q_1 to q_2 is a map $g : I \to T$, such that $g(0) = q_1$ and $g(1) = q_2$. T is *path-connected* if, given arbitrary points $q_1, q_2 \in T$, there is a path in T from q_1 to q_2.

Let $S^1 = \{(x,y) \in \Re^2 : x^2 + y^2 = 1\}$ be the circle of unit radius. Then, a *closed curve* (or *cycle*) in T is a map $f : S^1 \to T$. Choose a *base-point* $p \in S^1$, and consider it fixed for the rest of this discussion. Then, for a point $q \in T$, the set of (base-point preserving) cycles through q consists of maps $f : S^1 \to T$, such that $f(p) = q$.

The product of two cycles f and g through q is the cycle $f \circ g$ that is defined as the concatenation of f followed by g. The inverse f^{-1} of a cycle f through q is obtained by reversing the orientation of the map $f : S^1 \to T$. Two cycles through q are *equivalent* if there is a homotopy between them that fixes p. The set of equivalence classes of cycles through q forms a group under the induced product and inverse operations. This group is the *fundamental group* of T at q, and denoted $\pi(T,q)$.

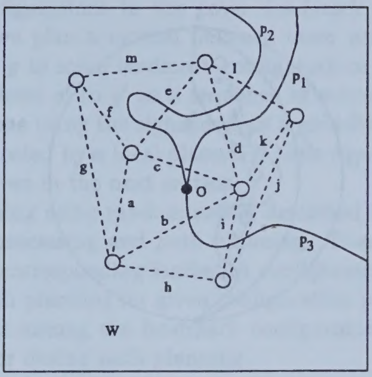


Figure 1: A flexible robot arm in a cluttered environment.

2.1.2 Group Theory

Given a set of symbols X, let X^{-1} denote the set of symbols $\{a^{-1} : a \in X\}$. A *letter* is an element of $X \cup X^{-1}$. A *word* w on X is a finite sequence $a_1 \dots a_k, k \geq 0$, of letters. The *length* of w, denoted $|w|$, is k. An *elementary transformation* of a word w consists of inserting or deleting a subword of the form aa^{-1} or $a^{-1}a$.

The set of all words on X is denoted $W(X)$. The equivalence relation $\sim$ on $W(X)$ is defined by $w_1 \sim w_2$ if w_2 can be derived from w_1 by a finite sequence of elementary transformations. The *free group* $F(X)$ on X is the set $W(X)/\sim$ of equivalence classes modulo $\sim$, endowed with the binary operation induced by concatenation. The *unit* of $F(X)$ is the equivalence class of the empty word ϵ, and the *inverse* of the equivalence class of $a_1 \dots a_k$ is the equivalence class of $a_k^{-1} \dots a_1^{-1}$. We shall, henceforth, always identify a word $w \in W(X)$ with its equivalence class in $F(X)$, and denote the (group) inverse of w by w^{-1}.

2.2 Topology-Guided Approach

To motivate our approach consider the flexible robot arm shown in Figure 1. There are seven obstacles in the workspace, which for simplicity are assumed to be points. The arm is attached to a fixed base O at one end. The other end is always assumed to touch the bounding square B. Thus the arm represents a path from O to B. Consider homotopy classes of *such* paths. In other words, two paths are homotopic if and only if one can be continuously transformed to the other, keeping one end fixed at O and the other moving always on B. For example, in Figure 1, paths p_1 and p_2 are homotopic, but p_1 and p_3 are not.

Observe now that the homotopy classes of such paths may be considered a free group on six generators. This may be seen by first noting that W with B identified to a point is homeomorphic to a sphere S with seven punctures, i.e., $S = \mathbf{S}^2 - \cup_{i=1}^{7}\{D_i\}$, where the D_i are disjoint disks on $\mathbf{S}^2$. See Figure 2. Now a path p in W joining O and B corresponds to a path p_S in S joining the north pole n and the south pole s. Draw a fixed path p_0 from s to n. Consider the correspondence $p \mapsto p_S \circ p_0$. This gives a (non-canonical, as it

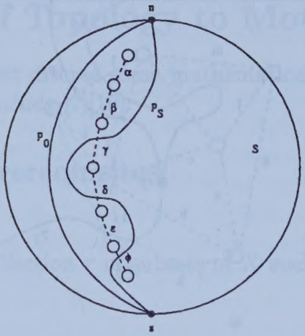


Figure 2: Paths on a punctured sphere.

depends on the choice of p_0) homotopy preserving correspondence between paths in W from O to B, and the loops on S based at n.

Observe, now that the fundamental group $\pi(S, n)$ is a free group on six generators. Geometrically, this may be seen by labeling the homotopy class of a loop at n in S by the sequence of edges from $\alpha, \beta, \gamma, \delta, \epsilon, \phi$, that it intersects, taking into account orientation. See Figure 2. The labeling then identifies the homotopy class of a loop with an element of the free group on six generators $\alpha, \beta, \gamma, \delta, \epsilon, \phi$, call this free group F_6. For example, the labeling of $p_S \circ p_0$ is $\gamma\delta^{-1}\phi$.

From this we conclude that there is indeed a (non-canonical) correspondence between the homotopy classes of paths in W from O to B and the free group F_6.

However, since we are dealing with the real workspace W, we prefer for our application to label homotopy classes of paths from O to B using the labels *a*–*m* of the 13 edges belonging to a triangulation of the convex hull of the point obstacles, as indicated in Figure 1. Specifically, the label of a path originating at O is the sequence of labels of the triangulation edges it intersects as we move along the path from O. For example, the label of path p_1 is *cdl*, and that of path p_2 is *ceedl*. Such labels are called the *signature* of a path. Clearly not all sequences of edge labels are valid signatures. For example, in Figure 1 *adkj* is not a valid signature as a path cannot intersect *d* right after intersecting *a*.

Note that paths from O to B belonging to the same homotopy class can have distinct signatures, e.g., p_1 and p_2 in Figure 1 belong to the same homotopy class but their signatures are not identical. Note also that some signatures, such as *ce*, represent a path from O that does *not* reach B.

We expect that a selection of configurations of the manipulator corresponding to every possible valid signature (an infinite set!) will provide a "complete cover" of the free configuration space C_f in the sense that an arbitrary configuration will almost certainly connect to some element of the cover using a simple local planner. Of course, we cannot implement an infinite cover. Further, configurations having "long" signatures, e.g., *cdkibafedl*, are unlikely to be attainable in practice since our manipulator in reality will have a few links of finite length.

Our heuristic is therefore simply to fix a set S of "short" valid signatures and (attempt to) generate a configuration of the manipulator for each signature in S, to obtain a cover of C_f. We call the configurations in the cover *landmark configurations.* Given two configurations s and g, we plan a motion between them using a simple local planner, by first connecting s and g to some landmark configurations s' and g', respectively, and then planning a motion from s' to g' as a sequence of successively reachable landmark configurations. This is done using the signatures as a guideline—two configurations may be more likely to be connected by a local planner if their signatures have a long common prefix. The details are given in the next section.

As with motion planning using randomization described in [5][8], we accomplish the task in two phases—preprocessing and path planning. Triangulation of the workspace and generation of $\mathcal{S}$ and corresponding landmark configurations are done in the preprocessing phase. Actual path planning for given configuration pairs is done in the planning phase. Connectivity tests among the landmark configurations could be done either in the preprocessing phase or during path planning.

The next section gives the details of our experiment.

3 Implementation and Experimental Results

3.1 A 7-DOF Manipulator

Our manipulator is a 7-DOF fixed-base robot arm consisting of seven equally long rigid links connected by six revolute joints. The joints are J_0, at the fixed base, to the outermost grasping end J_7. The links going outwards from J_0 are $L_1, \ldots, L_7$. The configuration of the robot arm is specified by the angles at $J_0, \ldots, J_6$, as a 7-tuple $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7)$. We do not allow configurations where two links intersect.

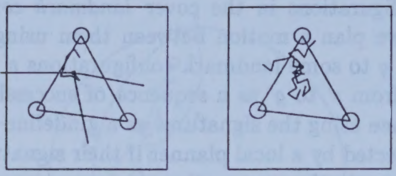
Given two configurations a and b, our local planner LP attempts to move the manipulator from a to b as a straight line motion between a and b in the 7-dimensional configuration space, by simultaneously rotating links L_i about joints J_{i-1}, each at some constant angular velocity. If this motion is collision-free and non-self-intersecting, then the local planner is successful in planning motion from a to b; otherwise it is deemed to have failed.

3.2 Preprocessing

For simplicity we assume that all obstacles in the workspace are circular. We experiment with six difference workspaces shown in Figures 3–8. Scene 6 is similar to a scene described in [5].

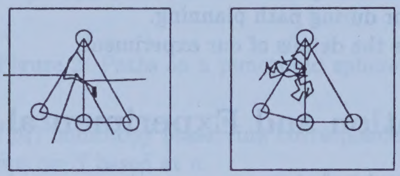
The preprocessing phase starts by triangulating the convex hull of the obstacles and labeling the triangulation edges, as shown in Figure 1. Generation of a set $\mathcal{S}$ of valid signatures is done automatically, based on some user specified parameters such as the maximum signature length.

Since our 7-DOF manipulator is not the fully flexible arm used in 2.2, there may not exist a configuration of the manipulator having a given valid signature — all corresponding configurations might be invalid, or the manipulator may not be long enough to cross all the triangle edges of the signature. So for each signature σ in $\mathcal{S}$, we *attempt* to generate a landmark configuration with signature σ using the following *midpoint heuristic*: Given a signature $\sigma = k_0k_1k_2 \ldots k_n$, we attempt to place the manipulator so that it first extends from the base through the midpoint of the edge labeled k_0, then through the midpoint of the edge labeled k_1, and so on.



(a) (b)

Figure 3: Scene 1: $n = 4$ (a) Topology-guided. (b) Random.



(a) (b)

Figure 4: Scene 2: $n = 6$ (a) Topology-guided. (b) Random.

3.3 Motion Planning

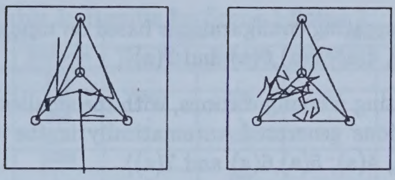
As we mentioned above, given two configurations s and g, we first connect them to some landmark configurations s' and g' and then plan a motion from s' to g' as a sequence of successively reachable landmark configurations. Configurations s' and g' are found from among those landmark configurations having signatures "similar" to those of s and g, respectively, using some heuristics.

The signatures of all landmark configurations can be thought of as being organized in a rooted ternary tree in which every node has the label of a triangulation edge it represents (the root is labeled λ), so that for every node v in the tree, the sequence of labels along the simple path from the root to v is a valid signature of some landmark configuration.

Given two landmark configurations c and c', we search for a motion between them as follows: Let u and u' be the vertices of the tree representing the signatures of c and c', respectively, and let $u = u_0, u_1, \ldots, u_m = u'$ be the simple path in the tree between u and u'. For each $1 \leq i \leq m-1$, let c_i be a landmark configuration whose signature is represented by node u_i. We then attempt to generate a motion from c to c' by concatenating motions from c to c_1, from c_1 to c_2, $\ldots$, and from c_{m-1} to c', using LP. Since this may not always succeed, we also attempt to connect a landmark configuration to additional landmark configurations using a heuristic whose details we do not give here.

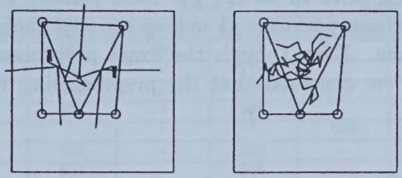
3.4 Results

We compared the performance of the topology-guided scheme with a random scheme. We did this by forming covers in two different ways:



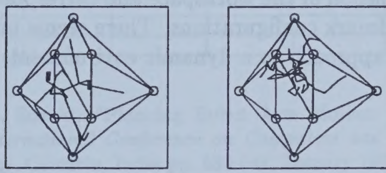
(a) (b)

Figure 5: Scene 3: $n = 6$ (a) Topology-guided. (b) Random.



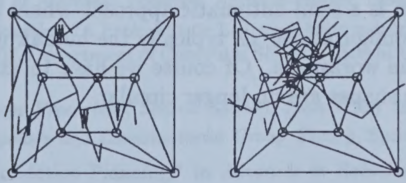
(a) (b)

Figure 6: Scene 4: $n = 8$ (a) Topology-guided. (b) Random.



(a) (b)

Figure 7: Scene 5: $n = 8$ (a) Topology-guided. (b) Random.



(a) (b)

Figure 8: Scene 6: $n = 15$ (a) Topology-guided. (b) Random.

(a) Automatically generating configurations based on topology as previously described (e.g. Figures 3(a), 4(a), 5(a),6(a) and 7(a)).

(b) Randomly generating n configurations, with the smallest value of n being the number of configurations generated automatically in the topology-guided approach. (e.g. Figures 3(a), 4(a), 5(a),6(a) and 7(a)).

We compared the number of motions that could be planned out of 1000 pairs of randomly generated source-goal pairs using the local planner LP for both approaches.

Tables 1–2 show the sizes n of the covers and the corresponding numbers of paths planned (out of 1000) for randomly generated source-goal pairs. For path planning using randomization, we used a threshold distance in the configuration space within which two configurations would have to lie for the local planner to be applied. As we see, the topology based approach performs as well as the approach based on "blind" random probing of the workspace. Even though the exact preprocessing times have not been reported in this paper, we observed that the preprocessing times for both approaches were comparable.

3.5 Observations

1. It is clear from the data that the topology-guided approach leads to smaller configuration graphs than the randomization approach for comparable performance.
2. We believe that our approach can be made adaptable to a dynamic environment containing moving obstacles. As the obstacles move, we can maintain and locally update the triangulation of the workspace and correspondingly update the set S of signatures and landmark configurations. There seems no evident method to adapt the randomization approach to a dynamic environment.

4 Conclusion

In this paper we have proposed a topology-based scheme for robot motion planning. In particular, we use the geometry of the workspace to determine a good cover of configurations based on selecting members of this cover from different homotopy classes. To verify this scheme we ran experiments using a multi-link robot arm in a cluttered environment. The results are promising as our data indicates. There is scope for further development along this approach. One example is implementing our approach for a dynamic environment. Another direction is a semi-automatic approach where the user intervenes in the selection of landmark configurations, as typically the human user has more insight into the complex parts of the workspace. Of course we have to extend these ideas to more general obstacles whose shapes are no longer circular.

References

[1] S. Guha, D. Suri and I. Suzuki, "Planning safe motion for 2- and 3-DOF manipulators by computing an approximate medial axis of the free configuration space," in *Proc. 5th IEEE International Conference on Emerging Technologies and Factory Automation*, Kauai, Hawaii, Vol. 2, pp. 467–474, November 1996.

[2] S. Guha, D. Suri and I. Suzuki, "Random probing to approximate medial axes and plan safe motion," in *Proc. 8th International Conference on Advanced Robotics (ICAR'97)*, Monterey, California, July 1997.

Table 1: Results for Scene 1-Scene 3

	Scene 1		Scene 2		Scene 3	
n	Topology	Random	Topology	Random	Topology	Random
4	1000	875	-	-	-	-
6	-	900	1000	725	996	728
7	-	906	-	-	-	-
8	-	906	-	804	-	730
9	-	1000	-	-	-	-
10	-	-	-	999	-	699
12	-	-	-	999	-	999
14	-	-	-	999	-	998
15	-	-	-	1000	-	-
16	-	-	-	-	-	999
18	-	-	-	-	-	1000

Table 2: Results for Scene 4-Scene 6

	Scene 4		Scene 5		Scene 6	
n	Topology	Random	Topology	Random	Topology	Random
8	977	882	694	514	-	-
10	-	927	-	778	-	-
12	-	984	-	778	-	-
14	-	975	-	989	-	-
15	-	-	-	-	585	402
16	-	984	-	989	-	-
18	-	993	-	996	-	-
20	-	991	-	996	-	610
25	-	982	-	993	-	605
30	-	993	-	1000	-	613
35	-	999	-	-	-	679
40	-	1000	-	-	-	810
45	-	-	-	-	-	771
50	-	-	-	-	-	770
60	-	-	-	-	-	799
70	-	-	-	-	-	814
80	-	-	-	-	-	787
90	-	-	-	-	-	791
100	-	-	-	-	-	795

[3] S. Guha, D. Suri and I. Suzuki, "Planning Robot Arm Motion: Applications of Topology," in *Proceedings of the International Conference on Computers and Devices for Communication (CODEC'98)*, Science City, Calcutta, India pp. 539-544, January 1998.

[4] Y. K. Hwang and N. Ahuja, "Gross Motion Planning - A Survey," *ACM Computing Surveys*, Vol. 24, No. 3, 1992, pp. 219–291.

[5] L. Kavraki and J.-C. Latombe, "Randomized Preprocessing of Configuration Space for Fast Path Planning," Technical Report STAN-CS-93-1490, Dept. of Comp. Sci., Stanford Univ., Stanford, CA, September 1993.

[6] J.-C. Latombe, *Robot Motion Planning*, Kluwer Academic Publishers, 1991.

[7] J. R. Munkres, *Elements of Algebraic Topology*, Addison-Wesley, 1984.

[8] M. H. Overmars and P. Švestka, "A Probabilistic Learning Approach to Motion Planning," Technical Report UU-CS-1994-03, Department of Computer Science, Utrecht University, The Netherlands, January 1994.

[9] W. K. Pratt, *Digital Image Processing, Second Edition*, John Wiley and Sons, New York, 1991.

[10] J. Stillwell, *Classical Topology and Combinatorial Group Theory*, Springer-Verlag, 1980.

[11] C. K. Yap, "Algorithmic Motion Planning," in *Advances in Robotics Vol. 1*, Schwartz and Yap Eds., Lawrence Erlbaum Pub., New Jersey, 1987, pp. 95–143.

Active Compensation of Flexibility in Belt-Driven Drive Trains

R. B. Gorbet* S. A. Bortoff†
Department of Electrical & Computer Engineering
University of Toronto
Toronto, Canada M5S 3G4

September 21, 1998

Abstract

This paper discusses a novel method to compensate the effects of flexibility in belt-driven drive trains. Such drives are found in robot manipulators, for example, where drive flexibility is undesirable. In some situations, bandwidth limitations can make flexibility difficult to compensate using the primary actuator. In others, the control for the primary actuator is pre-defined and cannot be easily altered. We have introduced a second actuator, called a *dancer*, which applies a force to the belt at a point between the pulleys. Using backstepping to control belt tension with the dancer, the system transfer function from input pulley to output pulley can be made exactly rigid. The performance of the dancer-actuated system compares favourably with a composite controller designed for the original belt drive.

1 Introduction

Flexible materials are often used to couple actuators to mechanical systems. For example, the links of a robotic manipulator may be driven by a flexible belt or cable, as illustrated in Figure 1. This allows the servos to be mounted at the stationary base, reducing manipulator mass and inertia. If the flexible dynamics are neglected in the design of a position or force controller, closed-loop performance may suffer, and instability can result. This observation has motivated much research into the compensation of flexible drive train dynamics [1, e.g.].

In Figure 1, J_i, q_i and r_i are the pulley inertia, angle and radius, respectively, for both output ($i = 1$) and input ($i = 2$). The equations of motion for the system are

$$J_1\ddot{q}_1 = r_1(T_b - T_t) \quad = \quad 2kr_1(r_2q_2 - r_1q_1), \tag{1}$$

$$J_2\ddot{q}_2 = r_2(T_t - T_b) + u_2 = -2kr_2(r_2q_2 - r_1q_1) + u_2. \tag{2}$$

---

[*] This research supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant OGP0138423, and the NSERC Postdoctoral Fellowship Program.

[†] Please direct correspondence to S. A. Bortoff, phone: (416)-978-0562, fax: (416)-978-0804, email: `bortoff@control.toronto.edu`.

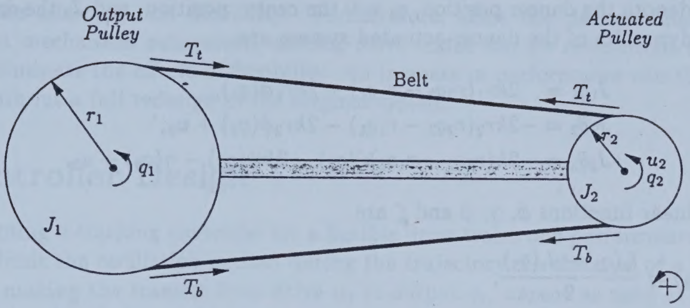


Figure 1: Belt-driven system.

If the spring constant k is large relative to J_1 and J_2, the system is *singularly perturbed.* Controllers for such systems are commonly designed using a time-scale separation technique known as composite control[2, e.g.]. If the time-scale separation is significant, the four-dimensional system (1)-(2) can be approximately decoupled into two smaller systems, one slow and one fast. Controls u_s and u_f are then designed separately for these two sub-systems, and added to form the overall control. This design will be carried out in more detail in Section 2.1.

Unfortunately, composite control has a disadvantage, especially in systems where inertias J_1 and/or J_2 are large: A single physical actuator is required to generate both u_s and u_f. If the inertias are large, u_s may contain large amplitude torque commands, requiring a large actuator. At the same time, u_f will contain high-frequency, often high-amplitude spectral components. The single actuator may not be able to satisfy both of these requirements simultaneously.

In [3], we introduced the idea of adding a second physical actuator called a *dancer*, which compensates the flexible dynamics in a belt-driven drive train by directly controlling belt tension. The dancer-actuated system is shown in Figure 2. The dancer is the link with the two small idler pulleys at either end. The dancer translates linearly, up and down as a unit, orthogonal to the horizontal link, and is actuated by a force u_3; the distance between the two idlers is fixed at $2d_3$. When the dancer moves up or down, it changes the relative tensions in the two belt spans. The rotational inertia of the two idlers is assumed to be zero, as is the mass of the belt, so that the belt tension is the same on the left-hand and right-hand sides of the dancer.

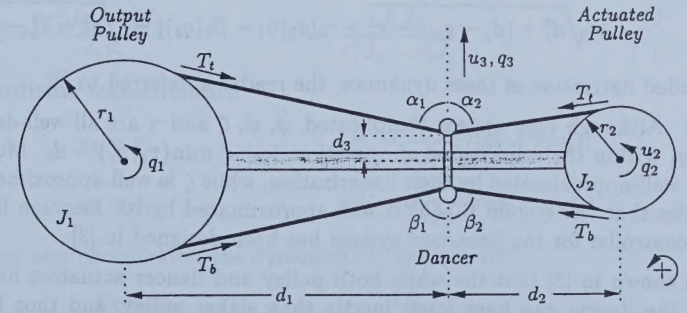


Figure 2: Belt-driven system with dancer.

Let q_3 denote the dancer position, $q_3 = 0$ the center position, and J_3 the dancer mass. Then the dynamics of the dancer-actuated system are

$$\begin{aligned} J_1\ddot{q}_1 &= 2kr_1(r_2q_2 - r_1q_1) + 2kr_1\phi(q_3), & (3)\\ J_2\ddot{q}_2 &= -2kr_2(r_2q_2 - r_1q_1) - 2kr_2\phi(q_3) + u_2, & (4)\\ J_3\ddot{q}_3 &= -2k(r_2q_2 - r_1q_1)\zeta(q_3) - 2k\psi(q_3) - \gamma(q_3) + u_3. & (5) \end{aligned}$$

The non-linear functions ϕ, γ, ψ and ζ are

$$\begin{aligned} \phi(q_3) &= \frac{l_b(q_3) - l_t(q_3)}{2},\\ \gamma(q_3) &= -T_0[\cos(\alpha_1(q_3)) + \cos(\alpha_2(q_3)) - \cos(\beta_1(q_3)) - \cos(\beta_2(q_3))],\\ \psi(q_3) &= \frac{l_b(q_3)[\cos(\beta_1(q_3)) + \cos(\beta_2(q_3))]}{2} - \frac{l_t(q_3)[\cos(\alpha_1(q_3)) + \cos(\alpha_2(q_3))]}{2},\\ \zeta(q_3) &= \frac{\cos(\alpha_1(q_3)) + \cos(\alpha_2(q_3)) + \cos(\beta_1(q_3)) + \cos(\beta_2(q_3))}{2}, \end{aligned}$$

where

$$\begin{aligned} \alpha_1(q_3) &= \arctan\left(\frac{d_3 + q_3}{d_1}\right) + \arctan\left(\frac{\sqrt{d_1^2 + (d_3 + q_3)^2 - r_1^2}}{r_1}\right),\\ \alpha_2(q_3) &= \arctan\left(\frac{d_3 + q_3}{d_2}\right) + \arctan\left(\frac{\sqrt{d_2^2 + (d_3 + q_3)^2 - r_2^2}}{r_2}\right),\\ \beta_1(q_3) &= \arctan\left(\frac{d_3 - q_3}{d_1}\right) + \arctan\left(\frac{\sqrt{d_1^2 + (d_3 - q_3)^2 - r_1^2}}{r_1}\right),\\ \beta_2(q_3) &= \arctan\left(\frac{d_3 - q_3}{d_2}\right) + \arctan\left(\frac{\sqrt{d_2^2 + (d_3 - q_3)^2 - r_2^2}}{r_2}\right), \end{aligned}$$

and

$$\begin{aligned} l_t(q_3) &= \sqrt{d_1^2 + (d_3 + q_3)^2 - r_1^2} + r_1[\alpha_1(0) - \alpha_1(q_3)] - \sqrt{d_1^2 + d_3^2 - r_1^2}\\ &\quad + \sqrt{d_2^2 + (d_3 + q_3)^2 - r_2^2} + r_2[\alpha_2(0) - \alpha_2(q_3)] - \sqrt{d_2^2 + d_3^2 - r_2^2},\\ l_b(q_3) &= \sqrt{d_1^2 + (d_3 - q_3)^2 - r_1^2} + r_1[\beta_1(0) - \beta_1(q_3)] - \sqrt{d_1^2 + d_3^2 - r_1^2}\\ &\quad + \sqrt{d_2^2 + (d_3 - q_3)^2 - r_2^2} + r_2[\beta_2(0) - \beta_2(q_3)] - \sqrt{d_2^2 + d_3^2 - r_2^2}. \end{aligned}$$

For a detailed derivation of these dynamics, the reader is referred to [3].

Remark: Although they appear complicated, ϕ, ψ, ζ and γ are all well-defined for all values of q_3 within the valid region of operation, $|q_3| < \min\{r_1, r_2\} - d_3$. Moreover, ϕ, γ and ψ are well-approximated by their linearization, while ζ is well-approximated by $\zeta(0)$. This implies that the system (3)-(5) is well-approximated by its Jacobian linearization. An LQR controller for the linearized system has been designed in [3].

It was shown in [3] that the while both pulley and dancer actuators have a similar DC gain, the dancer can have lower inertia than either pulley, and thus has a higher bandwidth. The dancer is hence better suited than the drive actuator for the role of canceling the effect of belt flexibility. Furthermore, since the dancer actuator is an independent mechanical subsystem, existing drive trains can be retro-fitted in order to reduce or eliminate the effects of flexibility. An increase in performance can therefore be achieved without a full redesign of the original system.

2 Controller Design

When designing a tracking controller for a flexible drive train, one performance objective may be to limit the oscillation excited during the trajectory. In the case of a belt drive, this means making the transfer from drive u_2 to output q_1, *appear* as rigid as possible.

In a rigid pulley system, the dynamics are

$$\frac{r_1^2 J_2 + r_2^2 J_1}{r_1 r_2}\ddot{q}_1 = u_2 \tag{6}$$

and the pulley angles are related by the algebraic constraint $r_1 q_1 = r_2 q_2$. This constraint defines a manifold in state space, called the *rigid manifold.* The magnitude of $r_1 q_1 - r_2 q_2$, the distance in state space from the rigid manifold, is therefore a measure of how "rigid" a system is.

In this section, we assume that, neglecting the flexible dynamics, a PD controller has already been designed for (6) and implemented on the flexible system (1)-(2):

$$u_2 = u_{pd} = \frac{r_1^2 J_2 + r_2^2 J_1}{r_1 r_2}[-k_p(q_1 - q_1^d) - k_d(\dot{q}_1 - \dot{q}_1^d) + \ddot{q}_1^d]. \tag{7}$$

The simulations of Figure 3 show that this control does a reasonable job of tracking a smoothed step. However, oscillations in the system response are apparent from the tracking error. The goal of the following controller designs is to augment the existing system so as to reduce or eliminate these oscillations. Detailed description and discussion of the simulations is given in Section 3.

2.1 Composite Control

In considering strategies to increase the apparent rigidity of the system, we first take a composite control approach. A fast corrective control is designed, to be added to the original PD control. The goal of the corrective control is to improve the attractiveness of the rigid manifold, thus reducing the oscillations observed in the system response.

We first write the dynamics (1)-(2) in singularly perturbed form. Define the constants

$$\bar{J} = \frac{r_1^2 J_2 + r_2^2 J_1}{r_1 r_2}, \qquad J^* = \frac{J_1 J_2}{r_1^2 J_2 + r_2^2 J_1}, \qquad \varepsilon = \sqrt{\frac{J^*}{2k}},$$

and the coordinate transformation

$$\begin{aligned} x_1 &= q_1, & \eta_1 &= \frac{2kr_1}{J_1}(r_2 q_2 - r_1 q_1), \\ x_2 &= \dot{q}_1, & \eta_2 &= \varepsilon\frac{2kr_1}{J_1}(r_2 \dot{q}_2 - r_1 \dot{q}_1). \end{aligned}$$

1000 In these new coordinates, the dynamics (1) and (2) are

$$\begin{aligned} \dot{x}_1 &= x_2, & \varepsilon\dot{\eta}_1 &= \eta_2, \\ \dot{x}_2 &= \eta_1, & \varepsilon\dot{\eta}_2 &= -\eta_1 + u_2/\bar{J}. \end{aligned}$$

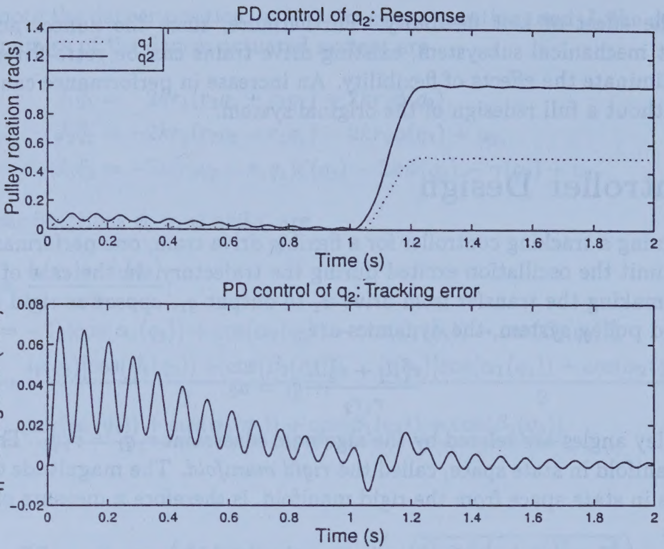


Figure 3: Original System Response with $u_2 = u_{pd}$.

If k is large relative to J_1 and J_2, then $\varepsilon \ll 1$, and there is significant time-scale separation between the "fast" η-subsystem and the "slow" x-subsystem.

Define a fast time-scale $\tau = t/\varepsilon$, and let x, η and u_2 be written in terms of slow $(\bar{\cdot})$ and fast $(\hat{\cdot})$ components: $x = \bar{x} + \hat{x}$, $\eta = \bar{\eta} + \hat{\eta}$, $u_2 = \bar{u}_2 + \hat{u}_2$. The slow components evolve in the slow time-scale t, while the faster components evolve in the "stretched" time-scale, τ. Note that the slow variable x actually has no fast component, so that $\hat{x} = 0$ and $x = \bar{x}$.

This time-scale separation allows the system to be approximately decoupled into two parallel systems, as in [2]. The slow system is

$$\dot{\bar{x}}_1 = \bar{x}_2, \tag{8}$$

$$\dot{\bar{x}}_2 = \bar{\eta}_1, \tag{9}$$

with the slow components of η defined by the algebraic equations

$$\bar{\eta}_1 = \bar{u}_2/\bar{J}, \tag{10}$$

$$\bar{\eta}_2 = 0. \tag{11}$$

Substituting (10) into (9) gives the rigid dynamics, (6). The fast system is described by

$$\frac{d\hat{\eta}_1}{d\tau} = \hat{\eta}_2,$$

$$\frac{d\hat{\eta}_2}{d\tau} = -\hat{\eta}_1 + \hat{u}_2/\bar{J}.$$

In composite control, we would now design two separate state-feedback controllers, $\bar{u}_2 = \bar{G}\bar{x}$ and $\hat{u}_2 = \hat{G}\hat{\eta}$, and combine the inputs, making use of (10) and (11), to form

$$u_2 = \bar{u}_2 + \hat{u}_2 = \bar{G}x + \hat{G}(\eta - \bar{\eta}). \tag{12}$$

In our case, the slow PD control has already been defined for the rigid system: $\bar{u}_2 = u_{pd}$. We wish to add to this control a fast corrective component which will improve the attractiveness of the rigid manifold, effectively the origin of the η-subsystem. A state feedback gain matrix $\hat{G}$ is computed to place the poles of the fast system at -20, five times faster than those of the slow PD control. The overall control law is then determined from (12) as

$$u_2 = u_{pd} + \hat{G}(\eta - [u_{pd}/\bar{J} \;\; 0]^T).$$

2.2 Backstepping Design

Rather than altering the drive pulley control, this section considers the addition of a dancer to the system. A backstepping controller is designed for the dancer, which not only improves the attractiveness of the rigid manifold but makes it invariant. For full details of the backstepping design technique, see [4, e.g.].

Recall that the system dynamics are given by (3)-(5). We assume $u_2 = u_{pd}$, and we wish to design a control for u_3 to force $r_1 q_1 = r_2 q_2$, so that the system *appears* rigid from u_2 to q_1. We first define the "output" as $z_1 = r_2 q_2 - r_1 q_1$, which we wish to force to zero. For this coordinate, choose the candidate Lyapunov function $V_1 = z_1^2/2$, which differentiated gives $\dot{V}_1 = z_1 \dot{z}_1$. We now define the first "virtual control": If $\dot{z}_1$ were a control input, then the control $\dot{z}_1 = -c_1 z_1$, $c_1 > 0$, ensures $\dot{V}_1$ is negative definite and $z_1 \to 0$. At this point, the design would be complete. However, $\dot{z}_1$ is not a control. Define instead a virtual control $v_1 = -c_1 z_1$, the error coordinate $z_2 = \dot{z}_1 - v_1$, and the new candidate Lyapunov function $V_2 = V_1 + z_2^2/2$. This is the beginning of the second iteration of the design.

Differentiating V_2 gives

$$\begin{aligned} \dot{V}_2 &= z_1 \dot{z}_1 + z_2 \dot{z}_2 \\ &= z_1(z_2 + v_1) + z_2(\ddot{z}_1 - \dot{v}_1) \\ &= -c_1 z_1^2 + z_2 \left[\left(1 - \frac{2k}{J^*}\right) z_1 - \frac{2k}{J^*}\phi(q_3) + \frac{r_2}{J_2} u_2 - \dot{v}_1 \right]. \end{aligned} \tag{13}$$

The control u_2 is pre-defined, so there is no input on the right-hand side of (13). Take $-\frac{2k}{J^*}\phi(q_3)$ as a "virtual" input, and define a second virtual control to make $\dot{V}_2$ negative definite:

$$v_2 = -\left(1 - \frac{2k}{J^*}\right) z_1 - \frac{r_2}{J_2} u_2 + \dot{v}_1 - c_2 z_2.$$

The new error coordinate is $z_3 = -\frac{2k}{J^*}\phi(q_3) - v_2$, and the new candidate Lyapunov function $V_3 = V_2 + z_3^2/2$.

Differentiating V_3, we have

$$\begin{aligned} \dot{V}_3 &= -c_1 z_1^2 - c_2 z_2^2 + z_3(z_2 + \dot{z}_3), \\ &= -c_1 z_1^2 - c_2 z_2^2 + z_3 \left[z_2 - \frac{2k}{J^*}\phi'(q_3)\dot{q}_3 - \dot{v}_2 \right], \end{aligned} \tag{14}$$

and once again there is no input in (14). Define a third virtual control

$$v_3 = -z_2 + \dot{v}_2 - c_3 z_3,$$

and the new error coordinate $z_4 = -\frac{2k}{J^*}\phi'(q_3)\dot{q}_3 - v_3$ and candidate Lyapunov function $V_4 = V_3 + z_4^2/2$.

Differentiating V_4 gives

$$\begin{aligned}\dot{V}_4 &= -c_1 z_1^2 - c_2 z_2^2 - c_3 z_3^2 + z_4(z_3 + \dot{z}_4),\\ &= -c_1 z_1^2 - c_2 z_2^2 .- c_3 z_3^2 + z_4 \left[z_3 - \frac{2k}{J^*}\phi''(q_3)\dot{q}_3^2 - \frac{2k}{J^*}\phi'(q_3)\ddot{q}_3 - \dot{v}_3 \right],\\ &= -c_1 z_1^2 - c_2 z_2^2 - c_3 z_3^2 \\ &\quad + z_4 \left[z_3 - \frac{2k}{J^*}\left(\phi''(q_3)\dot{q}_3^2 + \frac{\phi'(q_3)}{J_3}(-2k\zeta(q_3)z_1 - 2k\psi(q_3) - \gamma(q_3) + u_3) \right) - \dot{v}_3 \right].\end{aligned} \tag{15}$$

At this point, the dancer control u_3 has appeared in the right-hand side of (15), and we can define the dancer control law to ensure that $\dot{V}_4$ is negative definite:

$$u_3 = \gamma(q_3) + 2k(\psi(q_3) + \zeta(q_3)z_1) - \frac{J_3}{\phi'(q_3)}\phi''(q_3)\dot{q}_3^2 + \frac{J^* J_3}{2k\phi'(q_3)}(z_3 - \dot{v}_3 + c_4 z_4). \tag{16}$$

With u_3 defined as in (16), the zero dynamics, defined by the z-coordinates, are linear and fourth-order. Since $\dot{V}_4$ is negative definite, $z = 0$ is globally exponentially stable, and $z_1 \to 0$ as desired. In other words, the control (16) has the remarkable property that it forces the 6-dimensional system to the rigid manifold, so the input-output behavior from u_2 to q_1 is exactly second-order. The drawback is that the control u_2, which is pre-defined by (7), is differentiated twice during the backstepping algorithm, and so the desired trajectory must be sufficiently smooth. The backstepping control is also somewhat computationally complex.

3 Simulations & Discussion

The simulation paramters are shown in Table 1. In all cases, the initial condition[1] was $[q_1(0)\ q_2(0)\ q_3(0)\ \dot{q}_1(0)\ \dot{q}_2(0)\ \dot{q}_3(0)]^T = [0\ \ 0.1\ \ 0\ \ 0\ \ 0\ \ 0]^T$, which is off the rigid manifold and therefore excites the flexible dynamics. The reference trajectory is a smoothed polynomial unit step from $t = 1.0$ to $t = 1.25$ seconds.

Table 1: Parameters for the simulations, in MKS units.

Parameter	Value	Parameter	Value	Parameter	Value
r_1	0.2	J_1	1.0	d_1	1.0
r_2	0.1	J_2	0.5	d_2	0.335
k_p	16.0	J_3	0.1	d_3	0.05
k_d	8.0	T_0	100.0	k	50000.0

Figures 4 and 5 show the effect of composite control and backstepping control of the added dancer actuator, respectively. Note the two different scales for the composite control effort in Figure 4, and the separate scales for pulley angles and dancer displacement in Figure 5.

Both the composite control design and the backstepping design significantly improve the apparent rigidity of the system, as seen in the distance from the manifold and the

---

[1] q_3 and $\dot{q}_3$ are used only in the dancer-actuated system.

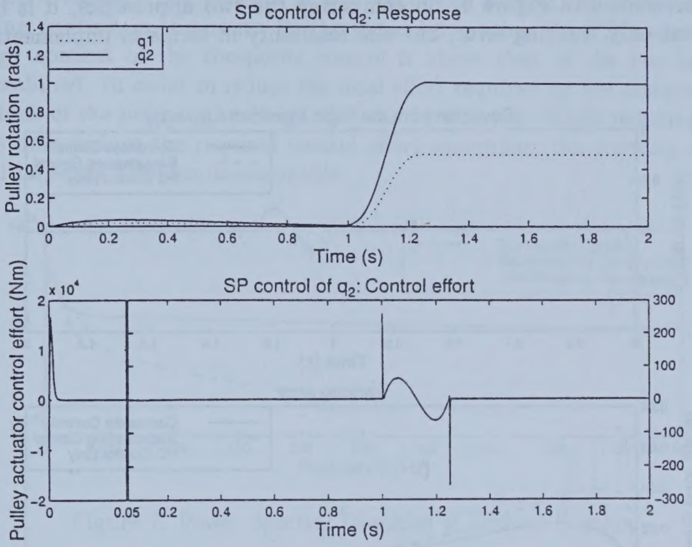


Figure 4: Original System Response: Composite Control.

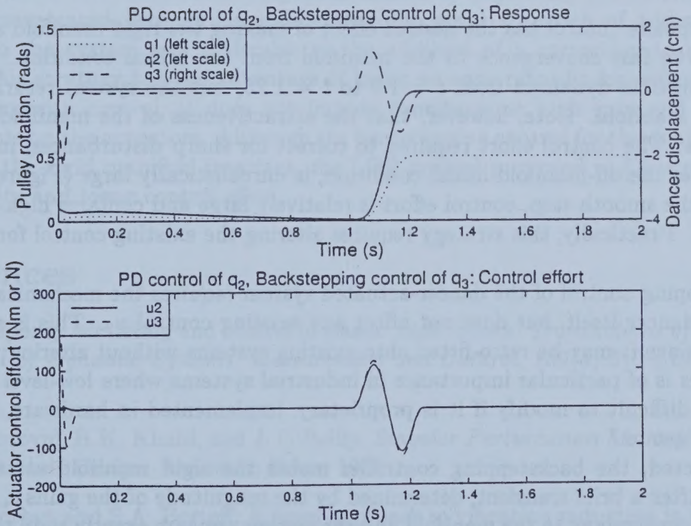


Figure 5: Dancer System Response: Backstepping Control.

tracking error, shown in Figure 6. In comparing the two approaches, it is important to consider not only tracking error, but also feasibility in terms of implementation and control effort.

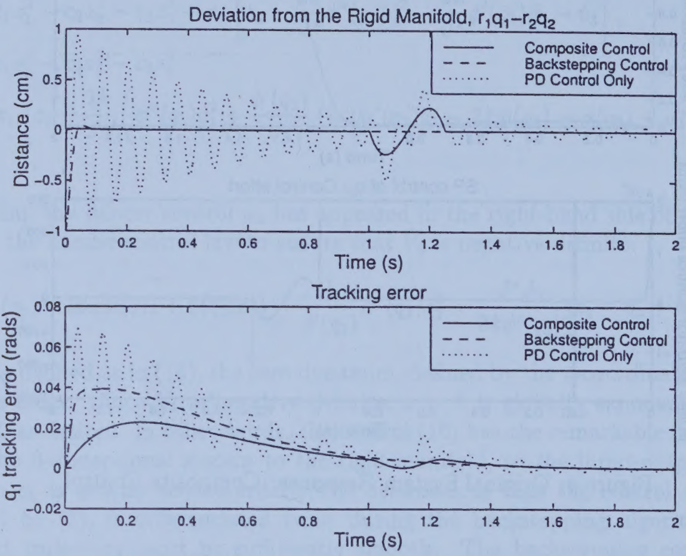


Figure 6: Comparison of Different Responses.

The composite control has the desired effect of making the rigid manifold attractive, as seen by the fast convergence to the manifold from the initial condition. The step excites the flexible dynamics from $t = 1.0$ to $t = 1.25$, but the system returns quickly to the rigid manifold. Note, however, that the attractiveness of the manifold comes at quite a price: The control effort required to correct for sharp disturbances, in this case in the form of the off-manifold initial condition, is unrealistically large (Figure 4). Even in tracking the smooth step, control effort is relatively large and contains high-frequency components. Practically, this strategy requires altering the existing control for the drive pulley, u_2.

Backstepping control of the dancer-actuated system requires the mechanical installation of the dancer itself, but does not affect any existing control u_2. This is seen as an advantage, since it may be retro-fitted onto existing systems without altering the pulley control. This is of particular importance in industrial systems where low-level joint control can be difficult to modify if it is proprietary, implemented in hardware, or simply difficult to "get to."

As expected, the backstepping controller makes the rigid manifold attractive and invariant. After a brief transient, determined by the magnitude of the gains c_i, the state evolution is constrained to the manifold and the system appears exactly rigid throughout the step tracking (Figure 6). The tracking error is only slightly worse than that obtained with composite control, and control effort is much more reasonable (Figure 5).

Spectral analysis of the control signals confirms that composite control requires higher control effort, at higher frequencies, and this from a single actuator. Figure 7 shows the power spectral densities (PSD) of the controls for the two approaches. The importance of the higher-frequency components in the composite control, as compared with the backstepping control, is apparent. Furthermore, integrating the PSDs shows that the total energy content of the composite control is above that of the two backstepping controls combined. In order to reduce the total effort required by the composite control, the magnitude of the poles in the design of u_2^f can be reduced. While reducing them from -20 to -8 does reduce the required control effort somewhat, the tracking error suffers and the initial effort remains unreasonable.

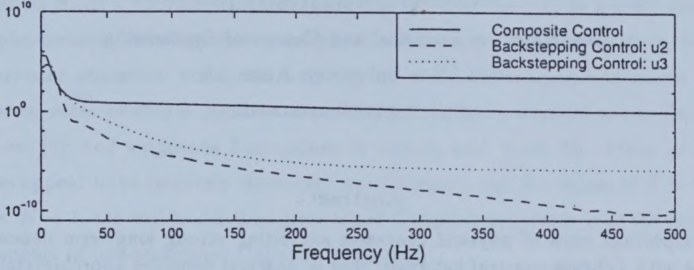


Figure 7: Power Spectral Densities of Control Signals.

4 Conclusions

This work investigated two strategies for improving the response of a flexible belt drive, in the case where a controller has already been defined by neglecting the flexible dynamics. It was demonstrated through simulation that the novel approach of adding a second actuator to the system was preferable to the addition of a correction to the existing control. This strategy has the advantage of being an easy retro-fit for existing systems. Unlike composite control, it does not impose simultaneous high gain and bandwidth requirements on the actuators. Although the backstepping control for the dancer succeeds in making the rigid manifold invariant, the LQR control proposed in [3] is preferble for its simplicity and lower control effort.

References

[1] M.W. Spong. Modeling and control of elastic joint robots. *Transactions of the ASME: Journal of Dynamic Systems, Measurement and Control*, 109(4):310–319, December 1987.

[2] P. Kokotović, H.K. Khalil, and J. O'Reilly. *Singular Perturbation Methods in Control: Analysis and Design.* Academic Press, 1986.

[3] R.B. Gorbet and S.A. Bortoff. A novel approach to vibration reduction in flexible belt drives. In D. Miller and L. Qiu, editors, *Topics in Control Theory and its Applications.* Springer-Verlag, to appear. Contact the authors for pre-prints.

[4] H.K. Khalil. *Nonlinear Systems.* Prentice Hall, 2nd edition, 1996.

Signal Detection in Fractional Gaussian Noise

Richard J. Barton
Veerendra Nauratanmull
Department of Electrical and Computer Engineering
Iowa State University, Ames, IA
email: barton@iastate.edu

Abstract

An important class of physical processes exhibiting strong long-term dependence are those with *1/f-type* spectral behavior; that is, spectral densities approximately proportional to f^{1-2H}, where f represents frequency and $1/2 < H < 3/2$ is a constant known as Hurst's Parameter. This type of behavior is observed in a great many different phenomena, including electro-magnetic radiation emitted by quasars; broadband underwater acoustical noise; the intensity of traffic on telecommunication networks; error rates on communication channels; frequency fluctuations in electrical oscillators; and amplitude fluctuations in speech and music. Many of these *1/f noises* appear to be relatively stationary and Gaussian and, for values of H in the range $1/2 < H < 1$, are well modeled as a fractional integral of white Gaussian noise, which is referred to as *fractional Gaussian noise.* In this paper we consider the problem of signal detection in fractional Gaussian noise. We briefly review the known results on this problem and present several extensions to these results which may have implications for wireless, spread-spectrum communication systems operating in heavily cluttered environments.

1 Introduction

In statistical signal processing applications, the lack of independence between observations has traditionally been handled by modeling the data as an autoregressive-moving-average (ARMA) process with relatively few parameters. Unfortunately, there are many physical processes that exhibit strong, positive, long-term correlations that are not well modeled by such an ARMA process. Such long-term dependence is often observed, for example, in geophysical data, where it takes the form of long periods of low or high values [3]. Similarly, on telecommunication channels, both error rates and traffic intensities display "bursty" behavior, in which a particular level of activity persists a relatively long period of time [8]. This tendency for low or high values to be followed by similar low or high values is indicative of a process possessing a long memory and is perhaps best characterized in terms of the spectral behavior of the process.

In particular, if the observed process is stationary, and the correlations between observations are positive and fall off so slowly that the covariance function is not integrable, then the spectral density of the process will be unbounded at the origin.

An important class of physical processes exhibiting strong long-term dependence are those with *1/f-type* spectral behavior; that is, spectral densities approximately proportional to f^{1-2H}, where f represents frequency and $1/2 < H < 3/2$ is a constant known as Hurst's Parameter. This type of behavior is observed in a great many different phenomena, including electro-magnetic radiation emitted by quasars; broadband underwater acoustical noise; the intensity of traffic on telecommunication networks [7]; error rates on communication channels [4]; frequency fluctuations in electrical oscillators [2]; and amplitude fluctuations in speech and music [6]. Many of these *1/f noises* appear to be relatively stationary and Gaussian and, for values of H in the range $1/2 < H < 1$, are well modeled as generalized Gaussian processes or as the (stationary) increments of non-stationary Gaussian processes.

In this paper, we consider the problem of signal detection in the presence of an additive stationary $1/f$ noise process. We restrict attention to values of H in the range $1/2 < H < 1$, and we adopt as our model of $1/f$ noise the class of so-called *fractional Gaussian noises* (FGN), which are the generalized derivatives of a class of self-similar Gaussian processes called *fractional Brownian motions* (FBM). We begin, in Section 2, by defining the models for FBM and FGN. In Section 3, we briefly review the known results for the problem of signal detection in FGN, which was first studied by Barton and Poor in [1]. In Section 4, we present some extensions to these results. The conclusions of this study and a discussion of future research directions are presented in Section 5.

2 Fractional Brownian Motion and Fractional Gaussian Noise

In this paper, we adopt the following definition of Fractional Brownian Motion. Let $\mathbf{B} \triangleq \{B(t); t \in \mathbf{R}\}$ be a standard Brownian motion and let $1/2 < H < 1$. The FBM process $B_H \triangleq \{B(t); t \in \mathbf{R}\}$ is defined by

$$B_H(t) \triangleq \frac{1}{\Gamma(H+\frac{1}{2})}\left[\int_{-\infty}^{0} (|t-\tau|^{H-\frac{1}{2}} - |\tau|^{H-\frac{1}{2}})dB(\tau) + \int_{0}^{t} |t-\tau|^{H-\frac{1}{2}} dB(\tau)\right] \qquad (1)$$

where $t \in \mathbf{R}$. Notice that if we extend our definition to include $H = 1/2$, we get

$$B_{1/2}(t) = B(t), \quad t \in \mathbf{R}.$$

It is straightforward to show that B_H is a zero-mean process with $B_H \equiv 0$ and covariance function is given by

$$K_{B_H}(s,t) = \frac{V_H}{2}\left(|s|^{2H} + |t|^{2H} - |t-s|^{2H}\right), \quad s,t \in \mathbf{R}, \qquad (2)$$

where

$$V_H \triangleq Var[B_H(1)] = \frac{-\Gamma(2-2H)\cos(\pi H)}{\pi H(2H-1)}. \tag{3}$$

Furthermore, the process has self-similar increments; that is

$$\{B_H(t_0+a\tau) - B_H(t_0); \tau \in \mathbf{R}\} \overset{d}{=} a^H \{B_H(\tau); \tau \in \mathbf{R}\} \tag{4}$$

From this it follows that, for any $\delta > 0$, the process $B_{H,\delta}$ given by

$$B_{H,\delta} \triangleq \frac{1}{\delta}\{B_H(t+\delta) - B_H(t); t \in \mathbf{R}\} \tag{5}$$

is a zero-mean stationary Gaussian process with covariance function

$$K_{B_{H,\delta}}(\tau) \cong V_H H(2H-1) \mid \tau \mid^{2H-2}, \quad \mid \tau \mid \gg \delta, \tag{6}$$

and power spectral density

$$\begin{aligned} S_{B_{H,\delta}}(\omega) &= \int_{-\infty}^{\infty} R_{B_{H,\delta}}(\tau) e^{-i\omega\tau} d\tau, && 0 \neq \omega \in \mathbf{R} \\ &\cong |\omega|^{1-2H}, && 0 < \mid \omega\delta \mid \ll 1. \end{aligned} \tag{7}$$

Hence the increments of FBM provide a simple model for processes with strong long-term dependence and $1/f$-type spectral behavior. The FGN process can be defined as the generalized derivative of FBM in the following manner. Notice that Equation (1) can be rewritten as

$$B_H(t) = \frac{1}{\Gamma(H-\frac{1}{2})} \int_0^t \int_{-\infty}^{\tau} \mid \tau - s \mid^{H-\frac{3}{2}} W(s)\, ds d\tau, \tag{8}$$

where $\{W(s)\}$ is the white-noise process corresponding to {B(t)}. If we formally define FGN as

$$W_H(t) \triangleq \frac{1}{\Gamma(H-\frac{1}{2})} \int_{-\infty}^{t} \mid t - \tau \mid^{H-\frac{3}{2}} W(\tau) d\tau, \tag{9}$$

then we get

$$B_H(t) \triangleq \int_0^t W_H(\tau)\, d\tau, \quad t \in \mathbf{R}, \tag{10}$$

which is the desired relationship. The FGN process is well defined as a generalized Gaussian process with covariance function

$$K_{W_H}(\tau) = V_H H(2H-1) \mid \tau \mid^{2H-2}, \quad \tau \in \mathbf{R}, \tag{11}$$

and power spectral density

$$S_{W_H}(\omega) = \mid \omega \mid^{1-2H}, \quad 0 \neq \omega \in \mathbf{R}. \tag{12}$$

3 Signal Detection in FGN

In this section, we present some useful results on the detection of deterministic signals in the presence of additive FGN, which were first derived in [1]. We consider a normalized version of the detection problem given by

$$\begin{aligned} H_0 : dY(t) &= \frac{\sigma}{\sqrt{V_H}} W_{H|T}(t)dt, & t \in [0,T], \\ H_1 : dY(t) &= s(t)dt + \frac{\sigma}{\sqrt{V_H}} W_{H|T}(t)dt, & t \in [0,T], \end{aligned} \tag{13}$$

where the normalizing factor $\sigma/\sqrt{V_H}$ reflects the assumption that the noise has "power" σ^2 (i.e., $Var\{Y(1)\} = \sigma^2$). The problem defined by Equation (13) will be non-singular if and only if for all $t \in [0,T]$,

$$s(t) = \frac{\sigma}{\sqrt{V_H}\Gamma(H-\frac{1}{2})} t^{H-\frac{1}{2}} \int_0^t (t-\tau)^{H-\frac{3}{2}} \tau^{\frac{1}{2}-H} \tilde{s}(\tau)\, d\tau, \tag{14}$$

where $\tilde{s} \in L^2([0,T])$, and for almost all $t \in [0,T]$,

$$\tilde{s}(t) = \frac{\sqrt{V_H}}{\sigma\Gamma(\frac{3}{2}-H)} t^{H-\frac{1}{2}} \frac{d}{dt}\left[\int_0^t (t-\tau)^{\frac{1}{2}-H} \tau^{\frac{1}{2}-H} s(\tau)\, d\tau\right]. \tag{15}$$

Assuming non-singularity, the log-likelihood ratio for Equation (13) is then given by

$$\log L(\mathbf{Y}) = <\mathbf{Y}, m>_{\mathbf{H(Y)}} - \frac{1}{2} <m,m>_{\mathbf{H(Y)}}, \tag{16}$$

where

$$m(t) \triangleq \int_0^t s(\tau)\, d\tau, \tag{17}$$

$$<m,m>_{\mathbf{H(Y)}} = \int_0^T [\tilde{s}(\tau)]^2\, d\tau, \tag{18}$$

and

$$<\mathbf{Y}, m>_{\mathbf{H(Y)}} = \frac{\sqrt{V_H}}{\sigma\Gamma(\frac{3}{2}-H)} \int_0^T \tilde{s}(t) t^{\frac{1}{2}-H} d_t \int_0^t (t-\tau)^{\frac{1}{2}-H} \tau^{\frac{1}{2}-H}\, d\mathbf{Y}(\tau). \tag{19}$$

It can be seen that $\log L(\mathbf{Y})$ is distributed in the following manner under the two hypotheses:

$$\begin{aligned} H_0 :\ \log L(\mathbf{Y}) &\sim N\left(-\frac{1}{2} <m,m>_{\mathbf{H(Y)}}, <m,m>_{\mathbf{H(Y)}}\right), \\ H_1 :\ \log L(\mathbf{Y}) &\sim N\left(\frac{1}{2} <m,m>_{\mathbf{H(Y)}}, <m,m>_{\mathbf{H(Y)}}\right). \end{aligned}$$

Assuming equal prior probabilities on H_0 and H_1, the minimum-probability-of-error detection rule takes the form

$$\delta(\mathbf{Y}) = \begin{cases} 1 & if \ \log L(\mathbf{Y}) \geq 0, \\ 0 & if \ \log L(\mathbf{Y}) < 0, \end{cases}$$

and the minimum probability of error is $\mathbf{P}_e^* = \Phi(-\frac{1}{2}\sqrt{<m,m>_{\mathbf{H(Y)}}})$, where Φ is the standard normal distribution function. Note that $\mathbf{P}_e^*$ is a decreasing function of the signal-to-noise ratio $<m,m>_{\mathbf{H(Y)}}$.

4 Detector Performance

4.1 Single sided Impulse Signal

Suppose that the signal s is of the form

$$s(t) = \frac{C\sqrt{2n+1}}{T^n}(T-t)^n, \quad t \in [0,T], \tag{20}$$

where C is any real constant and $n \geq 0, 1, 2, \ldots$. All of these signals have average power C^2 and satisfy Equation (14) with $\widetilde{s}$ given by (using Equation (15))

$$\widetilde{s}\,(t) = \sum_{i=0}^{n} \binom{n}{i} \frac{C\sqrt{V_H(2n+1)}\,t^{\frac{1}{2}+i-H}(2-2H+i)\Gamma(\frac{3}{2}-H+i)}{\sigma T^i \Gamma(3-2H+i)}$$

Hence, the SNR of the optimal detector is

$$\begin{aligned} <m,m>_{\mathbf{H(Y)}} &= \sum_{i=0}^{n}\sum_{k=0}^{n} \binom{n}{i}\binom{n}{k} \frac{C^2 V_H(2n+1)(2-2H+i)(2-2H+k)}{\sigma^2(2-2H+i+k)} \\ &\quad \frac{\Gamma(\frac{3}{2}-H+i)\Gamma(\frac{3}{2}-H+k)T^{2-2H}}{\Gamma(3-2H+i)\Gamma(3-2H+i)} \end{aligned} \tag{21}$$

For $H = \frac{1}{2}$, and $n = 0$ expression (21) reduces to

$$<m,m>_{\mathbf{H(Y)}} = \frac{C^2}{\sigma^2}T$$

as one would expect, since $H = \frac{1}{2}$ corresponds to white-noise.

Now suppose that s is of the form Equation (20) but that H is incorrectly identified; that is, suppose that $\log L(\mathbf{Y})$ is computed using Equations (15)-(19) on the basis of $H' \neq H$ where $\frac{1}{2} < H < 1$. In this case the log-likelihood ratio is distributed as follows:

$$\begin{aligned} H_0: &\quad \log L(\mathbf{Y}) \sim N(-\frac{1}{2}\mu, \nu^2), \\ H_1: &\quad \log L(\mathbf{Y}) \sim N(\frac{1}{2}\mu, \nu^2), \end{aligned}$$

where (using Equation (21))

$$\begin{aligned} \mu &= E\left\{ \frac{\sqrt{V_{H'}}}{\sigma\Gamma(\frac{3}{2}-H')} \int_0^T \widetilde{s}\,(t)t^{H'-\frac{1}{2}}d_t \int_0^t (t-\tau)^{\frac{1}{2}-H'}\tau^{\frac{1}{2}-H'}\,dY(\tau) \right\} \\ &= \int_0^T [\widetilde{s}\,(\tau)]^2\,d\tau \\ &= \sum_{i=0}^{n}\sum_{k=0}^{n} \binom{n}{i}\binom{n}{k} \frac{C^2 V_{H'}(2n+1)(2-2H'+i)(2-2H'+k)}{\sigma^2(2-2H'+i+k)} \\ &\quad \frac{\Gamma(\frac{3}{2}-H'+i)\Gamma(\frac{3}{2}-H'+k)T^2-2H}{\Gamma(3-2H'+i)\Gamma(3-2H'+i)}, \end{aligned} \tag{22}$$

and

$$\begin{aligned}\nu^2 &= Var\left\{\frac{\sqrt{V_{H'}}}{\sqrt{V_H}\sigma^2\Gamma(\frac{3}{2}-H')}\int_0^T \tilde{s}\,(t)t^{H'-\frac{1}{2}}d_t\int_0^t(t-\tau)^{\frac{1}{2}-H'}\tau^{\frac{1}{2}-H'}\,dY(\tau)\right\}\\ &= Var\left\{\sum_{i=0}^{n}\binom{n}{i}\frac{C\sqrt{V_H'(2n+1)}(2-2H'+i)\Gamma(\frac{3}{2}-H'+i)}{\sqrt{V_H}\sigma T^i\Gamma(3-2H'+i)\Gamma(\frac{3}{2}-H')}\right.\\ &\qquad\left.\int_0^T t^i d_t\int_0^t(t-\tau)^{\frac{1}{2}-H'}\tau^{\frac{1}{2}-H'}dY(\tau)\right\}\\ &= \sum_{i=0}^{n}\sum_{k=0}^{n}\binom{n}{i}\binom{n}{k}\frac{c^2{V_{H'}}^2(2n+1)(2-2H'+i)(2-2H'+k)}{[\sigma^2\Gamma(\frac{3}{2}-H')]^2}\\ &\quad\frac{\Gamma(\frac{3}{2}-H'+i)\Gamma(\frac{3}{2}-H'+i)H(2H-1)T^{2H-4H'+3}}{\Gamma(3-2H'+i)\Gamma(3-2H'+k)}E_n(H,H',i,k)\end{aligned}\tag{23}$$

Where $E_n(H,H',l,k)$ is given by

$$E_n(H,H',l,k)=\int_0^1\int_0^1\tau^{\frac{1}{2}-H'}(1-\tau)^{\frac{1}{2}-H'}v^{\frac{1}{2}-H'}(1-v)^{\frac{1}{2}-H'}|\tau-v|^{2H-2}$$
$$\left[1-l\sum_{i=0}^{l-1}\binom{l-1}{i}\frac{(1-\tau)^{l-i}\tau^i}{(l-i+\frac{1}{2}-H')}\right]\left[1-k\sum_{j=0}^{k-1}\binom{k-1}{j}\frac{(1-\tau)^{k-j}\tau^j}{(k-j+\frac{1}{2}-H')}\right]\tag{24}$$

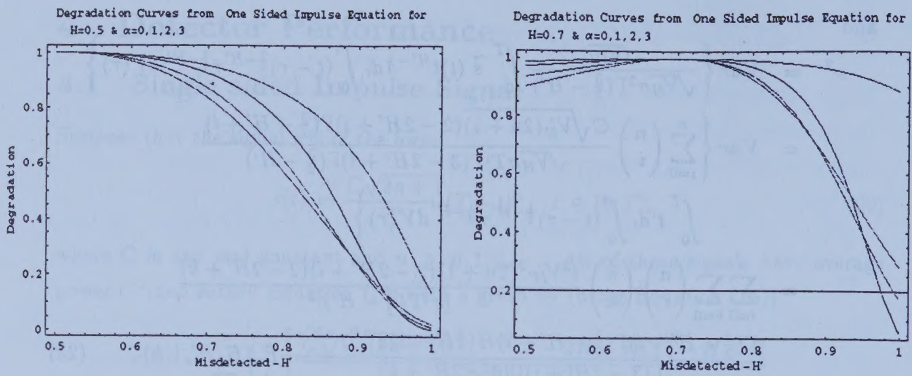
For any integer value of n the degradation factor can be obtained as

$$\delta_n(H,H')=\left[\frac{\sum_{i=0}^n\sum_{k=0}^n\binom{n}{i}\binom{n}{k}\frac{(2-2H'+i)(2-2H'+k)\Gamma(\frac{3}{2}-H'+i)\Gamma(\frac{3}{2}-H'+k)}{(2-2H'+i+k)\Gamma(3-2H'+i)(3-2H'+k)}}{\sum_{l_1=0}^n\sum_{k_1=0}^n\binom{n}{l_1}\binom{n}{k_1}\frac{(2-2H+l_1)(2-2H+k_1)\Gamma(\frac{3}{2}-H+l_1)\Gamma(\frac{3}{2}-H+k_1)}{(2-2H+l_1+k_1)\Gamma(3-2H+l_1)(3-2H+k_1)}}\right]^2\cdot$$
$$\frac{\sum_{l_2=0}^n\sum_{k_2=0}^n\binom{n}{l_2}\binom{n}{k_2}\frac{(2-2H+l_2)(2-2H+k_2)\Gamma(\frac{3}{2}-H+l_2)\Gamma(\frac{3}{2}-H+k_2)E_n(H,H,l_2,k_2)}{\Gamma(\frac{3}{2}-H)\Gamma(\frac{3}{2}-H)\Gamma(3-2H+l_2)(3-2H+k_2)}}{\sum_{l_3=0}^n\sum_{k_3=0}^n\binom{n}{l_3}\binom{n}{k_3}\frac{(2-2H'+l_3)(2-2H'+k_3)\Gamma(\frac{3}{2}-H'+l_3)\Gamma(\frac{3}{2}-H'+k_3)E_n(H,H',l_3,k_3)}{\Gamma(\frac{3}{2}-H')\Gamma(\frac{3}{2}-H')\Gamma(3-2H'+l_3)(3-2H'+k_3)}}$$

where $E_n(H,H',l,k)$ is as defined in equation(24) For $H=\frac{1}{2}$ the form of $\delta_n(H,H')$ is as above with

$$\begin{aligned}E_n(H,H',l,k)=\beta(2-2H',2-2H')-l\sum_{i=0}^{l-1}\binom{l-1}{i}\frac{\beta(2-2H+l-i,2-2H+i)}{(l-i+\frac{1}{2}-H')}\\ -k\sum_{j=0}^{k-1}\binom{k-1}{j}\frac{\beta(2-2H+k-j,2-2H+j)}{(l-j+\frac{1}{2}-H')}\\ +l\,k\sum_{i=0}^{l-1}\sum_{j=0}^{k-1}\binom{l-1}{i}\binom{k-1}{j}\frac{\beta(2-2H+l+k-i-j,2-2H+i+j)}{(l-i+\frac{1}{2}-H')(k-j+\frac{1}{2}-H')}\end{aligned}$$

Degradation factors for $n=0,1,2,3$ were evaluated numerically for various values of H and H' and are presented in Tables 1-4 and Figures 1-3. As these results indicate, in the constant signal case, there is little to be gained, and potentially much to be lost, by using a detector designed to match an assumed value of $H > 1/2$. The robust strategy is to use a simple integrate-and-dump detector. For the non-constant-signal case, there is much more to be gained by properly identifying the self-similarity parameter H.



(a) (b)

Figure 1: Degradation for Imp-Signal for (a) H = 0.5, and (b) H = 0.7

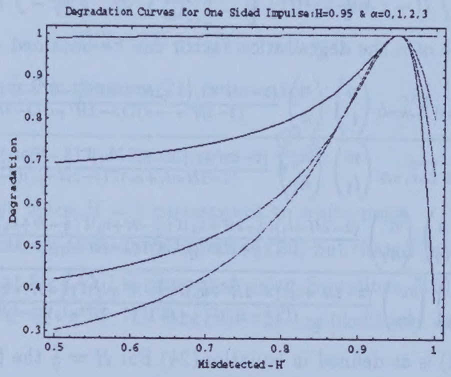


Figure 2: Degradation for Imp-Signal for H=0.95

Table 1: Degradation for Impulse Signal due to mismatch for $\alpha = 0$

$H \setminus H'$	0.500	0.600	0.700	0.800	0.900	0.950	0.975	0.999
0.500	1.0000	0.9910	0.9528	0.8534	0.6140	0.3823	0.2160	0.0098
0.600	0.9950	1.0000	0.9962	0.9622	0.8716	0.7850	0.7247	0.6533
0.700	0.9864	0.9960	1.0000	0.9934	0.9645	0.9341	0.9124	0.8862
0.800	0.9815	0.9902	0.9970	1.0000	0.9950	0.9870	0.9809	0.9734
0.900	0.9847	0.9898	0.9946	0.9984	1.0000	0.9994	0.9986	0.9972
0.950	0.9906	0.9932	0.9960	0.9982	0.9998	1.0000	1.0000	0.9996
0.975	0.9948	0.9962	0.9976	0.9988	0.9998	1.0000	1.0000	1.0000
0.999	0.9998	0.9998	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2: Degradation for Impulse Signal due to mismatch for $\alpha = 1$

$H \setminus H'$	0.500	0.600	0.700	0.800	0.900	0.950	0.975	0.999
0.500	1.0000	0.9869	0.9245	0.7392	0.3115	0.0866	0.0264	0.0042
0.600	0.9919	1.0000	0.9835	0.8892	0.5428	0.2364	0.1049	0.0275
0.700	0.9714	0.9898	1.0000	0.9706	0.7334	0.3946	0.1970	0.0576
0.800	0.9345	0.9569	0.9821	1.0000	0.8995	0.5963	0.3333	0.1041
0.900	0.8406	0.8618	0.8928	0.9419	1.0000	0.8837	0.6098	0.2136
0.950	0.6947	0.7121	0.7400	0.7923	0.9152	1.0000	0.8814	0.3701
0.975	0.5126	0.5253	0.5466	0.5892	0.7106	0.8915	1.0000	0.5572
0.999	0.0374	0.0383	0.0399	0.0433	0.0544	0.0805	0.1449	1.0000

Table 3: Degradation for Impulse Signal due to mismatch for $\alpha = 2$

$H \setminus H'$	0.500	0.600	0.700	0.800	0.900	0.950	0.975	0.999
0.500	1.0000	0.9768	0.8707	0.6127	0.2336	0.0840	0.0379	0.0135
0.600	0.9841	1.0000	0.9669	0.8047	0.4381	0.2324	0.1484	0.0860
0.700	0.9360	0.9766	1.0000	0.9404	0.6399	0.3915	0.2707	0.1716
0.800	0.8403	0.8933	0.9562	1.0000	0.8514	0.5950	0.4336	0.2850
0.900	0.6335	0.6827	0.7579	0.8773	1.0000	0.8839	0.7068	0.4841
0.950	0.4208	0.4559	0.5143	0.6265	0.8737	1.0000	0.9200	0.6702
0.975	0.2508	0.2724	0.3097	0.3871	0.6120	0.8920	1.0000	0.8132
0.999	0.0123	0.0134	0.0153	0.0196	0.0357	0.0810	0.2079	1.0000

Table 4: Degradation for Impulse Signal due to mismatch for $\alpha = 3$

$H \setminus H'$	0.500	0.600	0.700	0.800	0.900	0.950	0.975	0.999
0.500	1.0000	0.968-	0.8358	0.5691	0.2381	0.1026	0.0539	0.0234
0.600	0.9762	1.0000	0.9556	0.7736	0.4489	0.2810	0.2076	0.1469
0.700	0.9026	0.9654	1.0000	0.9287	0.6537	0.4605	0.3660	0.2827
0.800	0.7628	0.8448	0.9396	1.0000	0.8605	0.6655	0.5493	0.4377
0.900	0.5084	0.5796	0.6890	0.8553	1.0000	0.9126	0.7958	0.6506
0.950	0.3021	0.3489	0.4285	0.5816	0.8831	1.0000	0.9492	0.8021
0.975	0.1664	0.1933	0.2413	0.3439	0.6329	0.9194	1.0000	0.8969
0.999	0.0074	0.0086	0.0109	0.0163	0.0389	0.1086	0.2995	1.0000

4.2 DS-CDMA Signal

For the DS-CDMA signal of the form

$$s(t) = \sum_{n=0}^{N-1} C_n p(t - nT_c) \tag{25}$$

where $\{C_n\}$ represent a pseudo-noise spreading sequence with components in $\{-1, 1\}$. In this case, for $\frac{1}{2} < H < 1$, we get

$$\tilde{s}\,(t) = \frac{\sqrt{V_H}}{\sigma\Gamma(\frac{3}{2} - H)} t^{H-\frac{1}{2}} d_t \int_0^t (t-\tau)^{\frac{1}{2}-H} \tau^{\frac{1}{2}-H} \left[\sum_{n=0}^{N-1} C_n p(t - nT_c)\right] d\tau \tag{26}$$

For the interval $0 \leq t \leq T_c$ this reduces to

$$\tilde{s}_0\,(t) = \frac{C_0\sqrt{V_H} t^{\frac{1}{2}-H}(2-2H)\Gamma(\frac{3}{2} - H)}{\sigma\Gamma(3-2H)} \tag{27}$$

For t in the interval $nT_c \leq t \leq (n+1)T_c$, we get interval we get $\tilde{s}\,(t)$ to be

$$\begin{aligned} \tilde{s}_n\,(t) &= \frac{C_n\sqrt{V_H}}{\sigma\Gamma(\frac{3}{2} - H)} t^{\frac{1}{2}-H} \left[\left(\frac{nT_c}{t}\right)^{\frac{3}{2}-H} \left(1 - \frac{nT_c}{t}\right)^{\frac{1}{2}-H} + \right. \\ &\quad \left. (2-2H)\beta\left(\frac{3}{2} - H, \frac{3}{2} - H\right) \beta_{(1-\frac{nT_c}{t})}\left(\frac{3}{2} - H, \frac{3}{2} - H\right)\right] \\ &\cong \frac{C_n\sqrt{V_H} {T_c}^{\frac{1}{2}-H}}{\sigma\Gamma(\frac{3}{2} - H)} \alpha^{\frac{1}{2}-H} \end{aligned} \tag{28}$$

where we have substituted $t = (n+\alpha)T_c$ for $0 \leq \alpha \leq 1$ which defines the interval by $t = (n+\alpha)T_c$ Using the $\tilde{s}\,(t)$ obtained from (28) we get the mean and the variance as

$$\mu = \sum_{n=0}^{N-1} \mu_n, \qquad \nu^2 = \sum_{n=0}^{N-1} {\nu_n}^2 \tag{29}$$

where

$$\begin{aligned} \mu_n &= E\left\{\frac{\sqrt{V_{H'}}}{\sigma\Gamma(\frac{3}{2} - H')} \int_0^T \tilde{s}\,(t) t^{H'-\frac{1}{2}} d_t \int_0^t (t-\tau)^{\frac{1}{2}-H'} \tau^{\frac{1}{2}-H'}\, dY(\tau)\right\} \\ &= \frac{{C_n}^2 V'_H {T_c}^{2-2H'}}{(2-2H')[\Gamma(\frac{3}{2} - H')]^2\sigma^2} \end{aligned} \tag{30}$$

$$\begin{aligned} {\nu_n}^2 &= Var\left\{\frac{\sqrt{V_{H'}}}{\sqrt{V_H}\sigma\Gamma(\frac{3}{2} - H')} \int_0^T \tilde{s}\,(t) t^{H'-\frac{1}{2}} d_t \int_0^t (t-\tau)^{\frac{1}{2}-H'} \tau^{\frac{1}{2}-H'}\, dY(\tau)\right\} \\ &= \frac{{C_n}^2 {V_{H'}}^2 {T_c}^{2-4H'+2H} (n+1)^{2H'-1} H(2H-1)}{[\sigma\Gamma(\frac{3}{2} - H')]^4} E(H, H', n) \end{aligned} \tag{31}$$

where

$$\begin{aligned} E(H, H', n) = \int_0^1 \int_0^1 & ((1-\alpha)(n+\alpha)(1-\beta)(n+\beta))^{\frac{1}{2}-H'} \,|\, \alpha - \beta \,|^{-2+2H} \cdot \\ & \left(1 - \frac{(\frac{1}{2} - H')n(1-\alpha)}{(\frac{3}{2} - H')(1+n)} {}_2F_1[1, \frac{1}{2} + H', \frac{5}{2} - H', 1-\alpha]\right) \cdot \\ & \left(1 - \frac{(\frac{1}{2} - H')n(1-\beta)}{(\frac{3}{2} - H')(1+n)} {}_2F_1[1, \frac{1}{2} + H', \frac{5}{2} - H', 1-\beta]\right) d\alpha d\beta \end{aligned} \tag{32}$$

and ${}_2F_1[a, b, c, x]$ is the second-order hyper-geometric function.
For the case when $H = \frac{1}{2}$ $E(\frac{1}{2}, H', n)$ reduces to

$$E(\frac{1}{2}, H', n) = \int_0^1 (1-\alpha)^{1-2H'}(n+\alpha)^{1-2H'} \cdot \left[1 - \frac{(\frac{1}{2} - H')n(1-\alpha)}{(\frac{3}{2} - H')(1+n)} {}_2F_1[1, \frac{1}{2} + H', \frac{5}{2} - H', 1-\alpha]\right]^2 d\alpha \quad (33)$$

The degradation for an entire bit interval is then given by

$$\delta_N(H, H') = \left[\frac{2-2H}{2-2H'}\sqrt{\frac{\sum_{n=0}^{N-1} n^{2H-1} E(H, H, n)}{\sum_{m=0}^{N-1} m^{2H'-1} E(H, H', m)}}\right]^2 \qquad \frac{1}{2} < H < 1 \quad (34)$$

and for $H = \frac{1}{2}$ we have

$$\delta_N(\frac{1}{2}, H') = \left[\frac{1}{2-2H'}\sqrt{\frac{\sum_{n=0}^{N-1} E(\frac{1}{2}, \frac{1}{2}, n)}{\sum_{m=0}^{N-1} m^{2H'-1} E(\frac{1}{2}, H', m)}}\right]^2 \qquad \frac{1}{2} < H < 1 \quad (35)$$

where 'N' is the processing gain. These results can be easily extended for the case when the channel is white *i.e.* $H = \frac{1}{2}$. Numerical evaluation of this has not been completed at this time.

5 Conclusions and Future work

In this paper we have studied the problem of signal detection in fractional Gaussian noise. In particular, we have extended the earlier general results developed in [1] by evaluating optimal detector performance for a particular class of signals parameterized by relative bandwidth. In addition, we have quantified the degradation in receiver performance as a function of mismatch between the true and estimated values of the Hurst parameter for several particular values of the bandwidth parameter. Finally, we have demonstrated that the degradation in detector performance resulting from mismatch generally increases as the bandwidth of the transmitted signal increases. Our results imply that for broadband communication channels, proper identification of the Hurst parameter for the channel can be critical for achieving low bit-error rates.

References

[1] R. J. Barton and H. V. Poor, "Signal Detection in Fractional Gaussian Noise", *IEEE Transactions on Information Theory*, vol.34, No.5, September 1988.

[2] J. A. Barnes and D. W. Allen, "A statistical model of flicker noise," *Proceedings of the IEEE*, vol. 15, pp. 59-78, February 1966.

[3] H. Graf, "Long-Range Correlations and Estimation of the Self-Similarity Parameter," Ph.D. dissertation, Swiss Federal Institute of Technology, Zurich, 1983.

[4] B. B. Mandelbrot, "Self-similar error clusters in Communication Systems and the concept of conditional stationary," *IEEE Transactions on Communications*, vol.Com-13m [[. 71-90, march 1967.

[5] H. V. Poor, *An Introduction to Signal Detection and Estimation.* New York: Springer-Verlag, 1988

[6] R. F. Voss and J. Clarke, "1/f noise in music and speech," Nature, vol. 258, pp.317-318, November 1975.

[7] J. Beran, R. Sherman, M. S. Taqqu, W. Willinger,"Long-range dependence in variable bit rate traffic",*IEEE Transactions on Communications*, vol.43, pp.1566-1579, 1995.

[8] B. B. Mandelbrot, "Some noises with 1/f spectrum, a bridge between direct current and white noise," *IEEE Transactions on Information Theory*, vol. IT-13, pp. 289-298, April 1967.

A Dual System Approach To The Source Localization Problem

Mehmet E. Alpay
Electro-Scientific Industries
13900 NW Science Park Dr.
Portland, OR 97229
alpaym@elcsci.com

Molly H. Shor
Dept. of Electrical and Computer Engr.
Oregon State University
Corvallis, OR 97331
shor@ece.orst.edu

Abstract

In this paper, we develop a model-based approach for the localization of a point source with unknown and possibly time-varying strength for a class of distributed parameter systems. The technique involves the duality relationship between the primal and dual systems and the computation of the dual system's steady-state response. Two appropriately-placed point sensors are sufficient to locate the source for a one-dimensional problem. Three sensors are required for two-dimensional problems; four sensors for three-dimensional problems. The technique is applied to the two-dimensional heat conduction problem, and computer simulations are performed to test the method's performance in the presence of sensor noise.

1 INTRODUCTION

A distributed parameter system is an infinite-dimensional system whose behavior depends not only on time but also on spatial variables [1]. The mathematical model describing such a system usually consists of partial differential equations or integral equations. Solving for the dependent variable in such equations involves the computation of both its spatial and temporal variation. A finite-dimensional lumped parameter system approximation to a distributed parameter system can be obtained by "summarizing" (or "lumping") the infinite-dimensional state information into a finite number of state variables. Indeed, this is the approach taken in most numerical techniques developed for solving partial differential and integral equations [2].

Due to their higher degree of mathematical complexity, distributed parameter systems are harder to analyze and develop control schemes for than lumped parameter ones. However, many natural phenomena, such as electromagnetism, heat conduction, neutron transport and diffusion, are distributed parameter systems by their nature. Ane attempt to analyze accurately and regulate such phenomena requires the use of techniques developed specifically for distributed parameter systems. The recognition of this has prompted researchers in various fields of system theory to study control problems associated with distributed parameter systems and develop solution schemes (e.g., [3], [4]).

We define the source localization problem studied in this paper as the problem of finding the location of a single point source with possibly time-varying strength from the available output measurements for a class of distributed parameter systems. In this definition, the term "point source" refers to an input signal concentrated at a single point in the spatial domain of interest. The output measurements are obtained from a finite number of (point) sensors within the spatial domain. We assume *a priori* information about the mathematical model describing the distributed parameter system and use this model to develop our solution method. The precise mathematical set up for the source localization problem is given in §3.

The source localization problem has significant practical applications in environmental pollution monitoring. Consider, for example, the case of nuclear and chemical waste storage facilities. These facilities must be constantly monitored for accidental leaks and/or spills of hazardous contaminants. In the event of such an accident, it is crucial that the location of the leak/spill be quickly determined. Such a quick response would limit exposure of personnel, damage to the environment, and the cost of the subsequent clean-up effort (both in financial and environmental terms.) Since the dispersion as time evolves of a substance within space is a distributed parameter system, the solution to the associated source localization problem can be used for monitoring such storage facilities.

2 LITERATURE REVIEW

The source localization problem has been studied most extensively by researchers in the field of acoustics. These studies use the properties of sound-wave propagation in a medium with known acoustic characteristics. The methods developed for solving the acoustic source localization problem are mainly based on the concept of *spatial filtering.* The aim of the spatial filtering is to separate signals that have overlapping frequency content but originate from different spatial locations. A spatial filter is typically implemented by arranging sensors in a certain spatial configuration (usually in arrays), and taking a linear combination of delayed or filtered data from the sensors. there is *a priori* information about

Various types of spatial filters can be realized by changing the number and spatial configuration of the sensors, and the gains or filters associated with each sensor data channel. This process is called *beamforming* ([5]). The idea behind beamforming is to modify the sensor parameters to increase the sensor group's sensitivity to sounds coming from one direction, while suppressing the sounds coming from other directions. Different beamforming methods ([6]) and different ways of evaluating the data coming from the spatial filter ([7]) have been proposed and extensively tested. Their experiments in a water traditional array configuration. evaluation of the spatial filter output.

A quite different approach is taken in [8], in which the researchers use a sensor array to estimate the angles and relative arrival times of various multipath trajectories from the source. These trajectories are then back-propagated in time and space (numerically) to obtain an estimate of the source location.

The results obtained in the field of acoustic source localization depend heavily on the wave propagation characteristics, however, and are not readily applicable to the parabolic systems we study in this paper.

There are relatively few studies on the source localization problem applied to problems in other fields. In [9], researchers develop a neural networks optimization method for radioactive source localization. They assume *a priori* information about system parameters (physical dimensions, a transport model with known properties), and use a recurrent Hopfield neural network trained by a large number of simulated two-dimensional cases.

The source localization problem is studied for a more general setting in [10]. In this study, the researcher concentrates on parabolic systems, as we do in our research, and introduces a space of test functions that convert the solution of the source localization problem to the solution of a relevant system of algebraic equations. He establishes conditions under which the available output measurements are sufficient to identify uniquely the source location and finds the relationship between the (nonlinear) localization problem and the associated adjoint linear control problem.

The use of adjoint operators, dual systems, and the duality relationship for solving other *inverse problems* is common practice in many fields (e.g., [11]). Given the *direct problem* of finding the output of a system given a known input and a mathematical model describing the input/output relationship, there are two types of related inverse problems: 1) *causation*: given the (measured) output signal and a known system model, find the input signal, and 2) *identification*: given the input signal and (measured) resulting output signal, find (parameters of) the system model.

A common feature of inverse problems is their ill-posed nature; i.e., small changes in input data may give rise to large changes in the solutions ([12] and [13]). To circumvent this problem, one common approach is to convert the inverse problem to an optimization problem with an additional regularization term. The solution of the optimization problem itself often requires the construction of an appropriate dual system, and possibly applying the duality theorem (see [14] for the statement and proof of this theorem).

The source localization problem is an inverse problem of causation, while the problems studied in [11] is an identification problem. Since the admissible inputs for the source localization problem are highly restricted - single impulse in space - the ill-posedness of the general inverse problems is not a problem in our case. We do not convert the source localization problem to an optimization problem, and our use of the adjoint operators and dual systems is fundamentally different from [11] and other similar studies.

3 MATHEMATICAL BACKGROUND

Let Ω be an open bounded domain of n-dimensional Euclidian space $\Re^n$ with a boundary $\partial\Omega$. Consider the following problem for the linear parabolic equation

$$\frac{\partial u(\vec{r},t)}{\partial t} = Au(\vec{r},t) + v(t)\delta(\vec{r}-\vec{r}_{so}) \tag{1}$$

$$t \in T = (0,\theta), \quad \vec{r} \in \Omega \subset \Re^n, \quad Q = \Omega \times T, \quad \Sigma = \partial\Omega \times T$$

Initial condition : $u(\vec{r},0) = 0$.
Boundary conditions : $[a\,u(\vec{r},t) + b\,\partial u(\vec{r},t)/\partial\hat{n}]_\Sigma = 0$.

In the above, $v(t)$ is the unknown function for the source strength and $\vec{r}_{so}$ is the unknown source location. The operator A involves spatial derivatives of $u(\vec{r},t)$ up to second order. We assume that the system has a zero initial state, and we use a general *Robin's* boundary condition with $\hat{n}$ denoting the outward normal to the set Ω on the boundary $\partial\Omega$ of the set.

The output of the system in (1) consists of measurements of $u(\vec{r},t)$ at $M < \infty$ points in space, which yields

$$y(t) = \left[\ u(\vec{r}_{se_1},t) \quad u(\vec{r}_{se_2},t) \quad \cdots \quad u(\vec{r}_{se_M},t)\ \right], \tag{2}$$

where $\vec{r}_{se_1},\ldots,\vec{r}_{se_M}$ are the M point sensor locations.

The design parameters for the source localization problem are: 1) the number and locations of the sensors for output measurements, and 2) the location estimator - the evaluation algorithm that provides an estimate of the source location based on the available sensor data and the model of the system. Given these problem parameters, our design objectives are: 1) minimize the number of sensors, 2) optimize the sensor locations for robustness to sensor noise, and 3) minimize the amount of on-line computations.

4 THE DUAL SYSTEM METHOD

To develop the dual system method for the source localization problem, we first consider the system model (1) and (2) with a single sensor located at $\vec{r}_{se}$. The ideal (noise-free) measurements are given by the scalar function of time $y(t) = u(\vec{r}_{se}, t)$. We introduce a system dual to the one described by (1) and $y(t) = u(\vec{r}_{se}, t)$ as follows:

$$\frac{\partial z(\vec{r},t)}{\partial t} = -A^{*}z(\vec{r},t) - \lambda(t)\delta(\vec{r}-\vec{r}_{se}) \tag{3}$$

$$t \in T = (0,\theta), \quad \vec{r} \in \Omega \subset \Re^{n}, \quad Q = \Omega \times T, \quad \Sigma = \partial\Omega \times T$$

Final condition : $z(\vec{r},\theta) = 0$.

Here $\lambda(t)$ is the (arbitrary) source strength for the dual system and $\vec{r}_{se}$ is the sensor location for the primal system (the source location for the dual system). The operator A^{*} is the adjoint of the operator A of the primal system (1), and the dual system equations are to be solved reverse in time. The two operators, A and A^{*}, satisfy the following:

$$\int_{\Omega} g(Af)dV = \int_{\Omega} f(A^{*}g)dV \quad \forall f,g \in H_0^2(\Omega), \tag{4}$$

where $H_0^2(\Omega)$ is a Sobolev space with zero boundary conditions. We multiply (3) by $u(\vec{r},t)$ and integrate over space and time. By appropriate selection of boundary conditions for the dual system (3), and using (1), (4) and the initial and final conditions for $u(\vec{r},t)$ and $z(\vec{r},t)$, we obtain the equality:

$$\int_0^{\theta} v(t)z(\vec{r}_{so},t)dt = \int_0^{\theta} \lambda(t)y(t)dt. \tag{5}$$

Once $\lambda(t)$ is selected, the right-hand side of (5) can be evaluated numerically using the measurement $y(t)$. We wish to extract the source location information from the resulting integral equation. If we set $\lambda(t) = \lambda$ constant, then the dual system will converge (reverse in time) to a steady state $z_{ss}(\vec{r})$, which can be found explicitly by setting the time derivative in the dual system equation (3) to zero. We loosely define the *settling time* θ_s for the dual system as the time it takes for the dual system to get "sufficiently close" to this steady state, and we assume that $\theta_s \ll \theta$, where θ is the total observation time. Under these conditions, the integral equality of (5) can be approximated by

$$z_{ss}(\vec{r}_{so}) \int_0^{\theta} v(t)dt \cong \lambda \int_0^{\theta} y(t)dt, \tag{6}$$

where the function of source location is now outside the integral. However, with source strength $v(t)$ for the primal system unknown, this simplification still yields a single equation with two unknowns.

If we use not one but two sensors, then we can eliminate the dependence on $v(t)$. In the same manner as before, we construct a dual system in "two dimensions". The two states differ only in their respective source locations, $\vec{r}_{se_1}$ and $\vec{r}_{se_2}$. Setting the respective "source" strengths λ_i constant for each state, solving for their respective steady states $z_{ss1}(\vec{r})$ and $z_{ss2}(\vec{r})$, and defining $y_i(t) = u(\vec{r}_{se_i}, t)$, we obtain two equations

$$z_{ss1}(\vec{r}_{so}) \int_0^\theta v(t)dt \;\cong\; \lambda_1 \int_0^\theta y_1(t)dt \tag{7}$$

$$z_{ss2}(\vec{r}_{so}) \int_0^\theta v(t)dt \;\cong\; \lambda_2 \int_0^\theta y_2(t)dt, \tag{8}$$

where the variable $\vec{r}_{so}$ and the value of $\int_0^\theta v(t)dt$ are unknown. Dividing equation (7) by equation (8), we eliminate the dependence on the source strength $v(t)$ and obtain

$$\frac{z_{ss1}(\vec{r}_{so})}{z_{ss2}(\vec{r}_{so})} \cong \frac{\lambda_1 \int_0^\theta y_1(t)dt}{\lambda_2 \int_0^\theta y_2(t)dt}. \tag{9}$$

The only unknown variable is $\vec{r}_{so}$. Defining *localization function* $f_{12}(\vec{r}) = z_{ss1}(\vec{r})/z_{ss2}(\vec{r})$, we obtain an estimate $\hat{r}_{so}$ of the source location by solving

$$f_{12}(\hat{r}_{so}) = \frac{\lambda_1 \int_0^\theta y_1(t)dt}{\lambda_2 \int_0^\theta y_2(t)dt}. \tag{10}$$

The right-hand side of (10) is a number generated by filtering each output measurement $y_i(t)$, scaled by λ_i, $i = 1, 2$, through an integrator and taking the ratio. Solving (10) for $\vec{r}_{so}$ yields solution points for the one-dimensional case and contour lines and surfaces in the two- and three-dimensional cases.

4.1 Number of Sensors Required

It is evident from equation (10) that two sensors are sufficient to solve the source localization problem in a one-dimensional setting, provided that two sensor locations can be found so that the resulting localization function $f_{12}(\vec{r})$ is invertible within the spatial range of interest (Ω). For two and more dimensions, (10) is likely to have a set of solutions, and solving (10) indicates only on which contour lines or surfaces the source lies. The solution in such cases is to use more sensors. The minimum number of sensors required in a general n-dimensional setting is can be computed as follows: let us assume we have $M \geq 2$ sensors. Each sensor pair (i, j) corresponds to a localization function $f_{ij}(\vec{r})$. A unique estimate for the source location can be obtained if the set of equations

$$f_{ij}(\vec{r}) = \frac{\lambda_i \int_0^\theta y_i(t)\,dt}{\lambda_j \int_0^\theta y_j(t)\,dt} \qquad 1 \leq i, j \leq M, \quad i \neq j \tag{11}$$

has a unique solution $\hat{r}_{so}$ within the spatial domain Ω of interest. For an n-dimensional Ω, we need a minimum of n localization functions for this condition to hold. This imposes on the number of sensors M the requirement $C(M, 2) \geq n \Rightarrow M \geq (1 + \sqrt{8n+1})/2$, where $C(M, 2)$ represents the combination of M objects taken 2 at a time.

Nevertheless, if there are only a finite number of possible source sites known *a priori*, the minimum number of sensors required to solve the problem M_{min} is 2 in both the two- and three-dimensional cases. This reduction in the number of sensors is possible if there are two sensor locations such that the corresponding localization function does not have a contour that passes through more than one possible source site; i.e., the value of $f_{12}(\vec{r})$ at each potential source site completely identifies that site. This criterion forms the basis of strategic sensor placement in the two- and three-dimensional cases.

4.2 Robustness to Sensor Noise

The dual system method shows inherent robustness to sensor noise as a direct result of the particular way in which the sensor data is evaluated. Recalling (10), the source location is estimated by taking the ratio of "integrals" of the sensor data over the period θ of observation and constructing the contour of a pre-computed function (i.e., localization function) corresponding to this value. This low-pass filtering (through integration) during data evaluation suppresses the high-frequency noise effects on the method's performance.

4.3 Solving the Steady-State Dual System Equations

To implement the dual system method, it is necessary to compute the dual system's steady-state solution. This computation can be carried out in two ways: 1) Compute the adjoint operator A^* and then compute the dual system's steady-state solution analytically, or 2) Use the *finite element* method to compute the dual system's steady-state solution numerically by using the adjoint relationship (4) between A and A^*.

The finite element method only yields an approximate result. However, it is much simpler than the first method in cases where the adjoint operator A^* is not readily available, or the resulting dual system's steady-state equation is too complicated for existing analytical solution techniques. In this section, we review this technique and show how an approximation to the dual system's steady-state solution can be obtained without computing the adjoint operator A^*.

The steady-state equation for the dual system is given by

$$A^* z_{ss}(\vec{r}) + \lambda\delta(\vec{r} - \vec{r}_{se}) = 0 \tag{12}$$

with the boundary conditions to be determined so that the adjoint relationship (4) holds for all $u, z \in H^2(\Omega)$ that solve the primal (1) and dual (3) systems. Let $\{h_i(\vec{r})\}$, $i = 1, \ldots, N$, be a finite set of functions in $H^2(\Omega)$ that satisfy these boundary conditions. Define S as the span of these functions. An approximation of $z_{ss}(\vec{r})$ in S can be obtained as

$$\hat{z}_{ss}(\vec{r}) = \sum_{i=1}^{N} c_i\, h_i(\vec{r}) \tag{13}$$

where the c_i are undetermined coefficients. Substituting (13) into (12), we obtain

$$\begin{aligned} R(\vec{r}) &= A^* \hat{z}_{ss}(\vec{r}) + \lambda\delta(\vec{r} - \vec{r}_{se}) \\ &= \sum_{i=1}^{N} c_i\,(A^* h_i)(\vec{r}) + \lambda\delta(\vec{r} - \vec{r}_{se}) \end{aligned} \tag{14}$$

where $R(\vec{r})$ represents the error *residual*, which is due to $\hat{z}_{ss}(\vec{r})$ being an approximate solution. The objective of the finite element method is to find coefficients c_i that minimize this residual in some sense. The most commonly-employed minimization criterion is the *Method of Weighted Residuals* (MWR): the coefficients c_i are selected so that

$$\int_{\Omega} R(\vec{r}) w_j(\vec{r})\, dx = 0 \qquad j = 1, \ldots, N, \tag{15}$$

for a set of *weighting functions* $\{w_j(\vec{r})\}$. Substituting (14) into (15)

$$\sum_{i=1}^{N} c_i \int_{\Omega} (A^* h_i)(\vec{r}) w_j(\vec{r})\, dV + \lambda w_j(\vec{r}_{se}) = 0 \qquad j = 1, \ldots, N. \tag{16}$$

Using the adjoint relationship (4), (16) can be rewritten as

$$\sum_{i=1}^{N} c_i \int_{\Omega} (Aw_j)(\vec{r})h_i(\vec{r})\, dV + \lambda w_j(\vec{r}_{se}) = 0 \qquad j = 1, \ldots, N, \tag{17}$$

which can be expressed in compact matrix form as

$$\mathcal{A}c + w = \mathbf{0} \tag{18}$$

where

$$\mathcal{A} = \left[\int_{\Omega} (Aw_j)(\vec{r})h_i(\vec{r})\, dx\right]_{ji}, \quad c = [c_1 \cdots c_N]^T, \quad w = \lambda\, [w_1(\vec{r}_{se}) \cdots w_N(\vec{r}_{se})]^T \tag{19}$$

and $\mathbf{0}$ is a vector of zeros. The desired coefficients c_i can be found from

$$c = -\mathcal{A}^{-1}w \tag{20}$$

if $\mathcal{A}$ is invertible. Note that $\mathcal{A}$ can be computed without finding A^*.

There are numerous MWR techniques based on the general approach described above, for each particular choice of basis and weighting functions. One of the most common techniques is the *Galerkin Method*, in which the weighting functions are taken to be the same as the basis functions. In §5, we use the Galerkin Method with *Chapeau* basis functions to solve for the steady-state solution to the dual system of two-dimensional heat conduction equation with Robin's boundary conditions.

5 AN EXAMPLE - TWO-DIMENSIONAL HEAT CONDUCTION

The two-dimensional heat conduction equation is stated below:

$$\frac{\partial u(x,y,t)}{\partial t} = D_x \frac{\partial^2 u(x,y,t)}{\partial x^2} + D_y \frac{\partial^2 u(x,y,t)}{\partial y^2} + v(t)\delta(x - x_{so}, y - y_{so}) \tag{21}$$

$$0 < x, y < 1, \quad 0 < t < T$$

Initial condition : $u(x,y,0) = 0.$
Boundary conditions : $[u(x,y,t) + \partial u(x,y,t)/\partial \hat{n}]_{x,y=0} = 0.$

The state $u(x,y,t)$ is measured at three sensor locations to constitute the output

$$y(t) = \left[\; u(x_{se_1}, y_{se_1}, t) \quad u(x_{se_2}, y_{se_2}, t) \quad u(x_{se_3}, y_{se_3}, t) \;\right]^T. \tag{22}$$

In the above equations, (x_{so}, y_{so}) is the source location and (x_{se_k}, y_{se_k}), $k = 1, 2, 3$ are the (point) sensor locations. The vector $\hat{n}$ denotes the outward normal at the boundaries of the spatial region $(0,1) \times (0,1)$. The diffusion coefficients D_x and D_y in $x-$ and $y-$directions have values of 1.

The operator $A = D_x\,\partial/\partial x^2 + D_y\,\partial/\partial y^2$ is self-adjoint in $H_0^2((0,1) \times (0,1))$. Recalling (3), the equations for the system dual to (21) are given by

$$\frac{\partial z(x,y,t)}{\partial t} = -D_x \frac{\partial^2 z(x,y,t)}{\partial x^2} - D_y \frac{\partial^2 z(x,y,t)}{\partial y^2} - \lambda(t)\delta(x - x_{se}, y - y_{se}) \tag{23}$$

$$0 < x, y < 1, \quad 0 < t < T$$

Final condition : $z(x, y, 0) = 0$.
Boundary conditions : $[z(x, y, t) + \partial z(x, y, t)/\partial \hat{n}]_{x,y=0} = 0$.

To apply the dual system method to this problem, we consider a single source at (0.25,0.25) and place our three sensors at (0.1,0.3), (0.9,0.3) and (0.4,0.6). We use the analytical solution to the heat conduction equation (21) to generate the sensor data by truncating the series at a finite index $N = 10$. We first compute the steady-state solutions to the dual system, thence the localization functions, analytically (see [15]). We repeat these computations numerically using the finite element method. For the finite element method, we partition the $(0, 1) \times (0, 1)$ region by a 8×8 grid and construct the resulting forty-nine modified Chapeau basis functions satisfying the given boundary conditions. We then use the Galerkin MWR to solve for the unknown coefficients (see §4.3) and construct the dual system's approximate steady-state solutions, thence the approximate localization functions. The analytically- and numerically-computed versions of one of these localization functions are compared in Figures 1 and 2.

In our simulations we consider the following scenarios
1) Constant source strength, noise-free environment, perfect model match
2) Decaying source strength, noise-free environment, perfect model match
3) Constant source strength, noisy environment, perfect model match
4) Decaying source strength, noisy environment, perfect model match

In each of the four cases listed above, we evaluate the right-hand side of (9) for each sensor pair (i, j) to obtain an estimate α_{ij} of the value of the corresponding localization function at the source location (x_{so}, y_{so}). We then draw the contours of the analytical localization functions for these estimates. The intersection of these contours from all three localization functions yields an estimate of the source location. We repeat the process using the numerical (approximate) localization functions obtained from the finite element method. Our findings for all scenarios are summarized in Table 1.

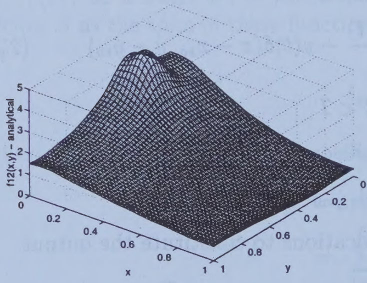


Figure 1: *Analytically-computed localization function* $f_{12}(x, y)$.

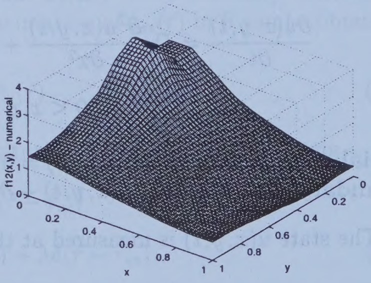


Figure 2: *Numerically-computed localization function* $f_{12}(x, y)$.

Some of the ideas developed in the paper are illustrated in Figures 3 and 4, using the data generated by computer simulations and the analytically-computed localization functions. In both figures, the star symbol represents the true source location (0.25, 0.25), and the filled circle and square represent the source location estimates for various cases.

It should be noted that the relationship between the random sensor noise and the accuracy of the eventual source location estimate is not straightforward. The dual system method does not provide an "exact" solution to the source localization problem even

				Analytical Localization Function		*Numerical Localization Function*	
Case	α_{12}	α_{13}	α_{23}	$(\hat{x}_{so}, \hat{y}_{so})$	$\lVert\text{error}\rVert_2$	$(\hat{x}_{so}, \hat{y}_{so})$	$\lVert\text{error}\rVert_2$
1	2.2930	1.5832	0.6905	(0.234,0.273)	0.0280	(0.229,0.242)	0.0225
2	2.1885	1.5336	0.7007	(0.247,0.255)	0.0058	(0.238,0.230)	0.0233
3	2.2069	1.6100	0.7295	(0.244,0.201)	0.0494	(0.226,0.173)	0.0807
4	2.3765	1.6041	0.6750	(0.225,0.298)	0.0541	(0.221,0.260)	0.0307

Table 1: *Simulation results*

under ideal conditions (noise-free environment, perfect model match) due the approximation in (9); i.e., the method starts off with an initial error *before* noise is introduced. There is no relation between the error resulting from the approximation in (9) and the additional errors introduced by sensor noise. When the magnitude of the additional errors introduced by sensor noise are comparable to the magnitude of the initial error, they may, in some instances, partially cancel it out, which improves the source location estimate accuracy. This interesting phenomenon is illustrated in Figure 3. To obtain this figure, the noise signals generated for Case 3 are scaled by factors ranging from 0 to 1, and outputs corrupted by various degrees of noise are computed. At each noise level, the α_{ij} are computed from the noisy outputs, and the corresponding localization function contours are drawn. As the noise level increases, the source location estimates initially improve but then get worse again. This behavior is understandable in light of the above discussion.

The localization function contours corresponding to a decaying-strength source ($v(t) = e^{-t}$) and a constant-strength source ($v(t) = 1$) are compared in Figure 4 in a noise-free environment with perfect model match. Clearly, the dual system method has performed much better for the decaying-strength source case. This somewhat unexpected result will be repeated for all such cases in which the source strength diminishes in time. The reason for this can be traced to the approximation

$$\int_0^\theta z(\vec{r}_{so}, t)v(t)dt \cong z_{ss}(\vec{r}_{so}) \int_0^\theta v(t)dt.$$

This approximation would be an exact equality if $z(\vec{r}_{so}, t)$ were to reach steady state (reverse in time) instantaneously. Still, the difference between the two quantities will be small if the integrand of the left-hand side is small during the transition period $t : \theta - \theta_s \rightarrow \theta$, which is the case for a decaying source strength function $v(t)$. Therefore, this interesting phenomenon is the rule, rather than the exception, for all decaying-strength sources.

6 CONCLUSION

In this paper, we have developed a technique for solving for the location of a point source with unknown and possibly time-varying strength in a parabolic distributed parameter system where the measurements are obtained via point sensors. The technique involves the computation of the steady-state distributions for the dual systems, which are then correlated with the sensor measurements via duality relationships to estimate the unknown source location for the primal system. We have developed a criterion for the selection of sensor locations and we have shown that two sensors meeting this criterion

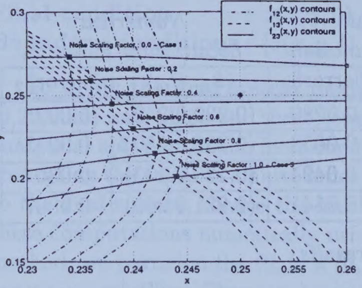


Figure 3: *Effects of sensor noise on the contours of the (analytically-computed) localization functions and the resulting source location estimates for Case 3.*

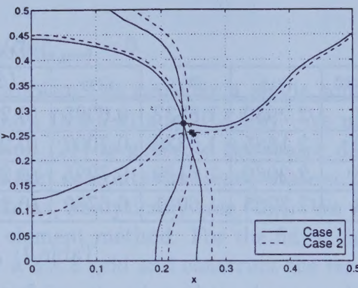


Figure 4: *Effects of decaying vs. constant source strength functions* ($v(t)$) *on the contours of the analytically-computed localization functions.*

are sufficient to solve for the source location for a one (spatial) dimensional problem, while more sensors are required for the two- and three-dimensional problems. Nevertheless, two sensors may be sufficient to solve the problem if there are only a finite number of potential source sites and two sensors yielding a localization function that does not have a contour passing through more than one potential source site can be found.

We have applied the dual system method to the two-dimensional heat conduction equation, and illustrated its performance under various conditions via computer simulations. In particular, we have examined the effects of sensor noise and source strength variation. We have argued, and demonstrated by simulations, that the relationship between the sensor noise level and the accuracy of the source location estimate is not straightforward, and that the method's performance might happen to improve as noise level *increases* in some cases. We have also observed that the dual system method performs better for cases involving decaying-strength sources. We have traced the roots of this phenomenon to the approximations made while developing the method, and we have argued that the same behavior can be expected in all cases with such sources.

The major weakness of the dual system method is that it is based on an approximation that introduces an initial error that is hard to compute and/or obtain an upper bound for. The strengths of the method include its ability to handle time-varying sources and its inherent robustness to the effects of zero-mean high frequency sensor noise. The method requires only a minimal number of sensors, and provides a criterion for the selection of the appropriate sensor locations. The amount and complexity of the on-line computations required for sensor data evaluation and source location estimation are within achievable limits. The bulk of the computations can be performed off-line with high degree of accuracy, which should further improve the dual system method's overall performance.

Possible topics for future studies on the dual system method include: 1) the analysis of the approximation (9) and how it affects the method's accuracy, 2) the use of a larger number of basis functions of higher order to construct better numerical approximations to the analytical (exact) localization functions, 3) improvement of the method's robustness to sensor noise through additional temporal filtering, and 4) an in-depth study of the principles laid out for appropriate sensor placement.

References

[1] K.L. Teo and N.U. Ahmed. *Optimal Control of Distributed Parameter Systems.* Elsevier North Holland, Inc., 1981.

[2] G. F. Pinder and L. Lapidus. *Numerical Solution of Partial Differential Equations in Science and Engineering.* John Wiley & Sons, 1982.

[3] A.G. Butkovskiy. *Distributed Control Systems.* Elsevier, 1969.

[4] A.K. Aziz. *Control Theory of Systems Governed by Partial Differential Equations.* Academic Press, 1977.

[5] K.M. Buckley and B.D. Van Veen. Beamforming: A versatile approach to spatial filtering. *IEEE Acous., Speech, Signal Proc. Mag.*, pages 4–24, 1988.

[6] R. Raspet, R. Hickling and W. Wei. Finding the direction of a sound source using a vector sound intensity probe. *J. Acous. Soc. Amer.*, 94:2408–2412, October 1993.

[7] H. Chuberre and J.-J. Fuchs. A deconvolution approach to source localization. *IEEE Trans. Signal Proc.*, 42:1462–1470, June 1994.

[8] I-T. Lu and P. Voltz. A time-domain backpropagating ray technique for source localization. *J. Acous. Soc. Amer.*, 95:805–812, February 1994.

[9] Y. Merlis, E. Wacholder and E. Elias. Artificial neural networks optimization method for radioactive source localization. *Nucl. Tech.*, 110:228–237, May 1995.

[10] A.Y. Khapalov. Location of unknown sources for parabolic systems on the basis of available observations. *Int. J. Syst. Sci.*, 25(8):1305–1322, August 1994.

[11] J.H. Renken. Use of adjoint transport solutions for inverse problems. *Trans. Am. Nucl. Soc.*, 56:256–257, 1988.

[12] A.N. Tikhonov. *Solutions of Ill-posed Problems.* Winston, 1977.

[13] C.W. Groetsch. *Inverse Problems in the Mathematical Sciences.* Vieweg, Braunschweig, Germany, 1993.

[14] D.G. Luenberger. *Optimization by Vector Space Methods.* John Wiley & Sons, 1969.

[15] M.E. Alpay. *Model-Based Solution Techniques for the Source Localization Problem in Distributed Parameter Systems.* PhD thesis, Oregon State University, July 1998.

A WDM Ring Network for Incremental Traffic *

Galen Sasaki
University of Hawaii
2540 Dole St, Honolulu, HI 96822
sasaki@spectra.eng.hawaii.edu

Ori Gerstel and Rajiv Ramaswami
Tellabs
15 Skyline Dr, Hawthorne, NY 10532
{ori, rajiv}@tellabs.com

Abstract: An optical *wavelength division multiplexed* (WDM) ring network is presented that supports dynamic (time-varying) traffic. The traffic streams are at lower rates (e.g., OC-3, 155Mb/s) than the lightpath capacities (e.g., OC-48, 2.5 Gb/s). The network has a nonblocking property if the streams are assumed to never terminate, i.e., *incremental traffic*.

A design algorithm is presented that minimizes the cost of the network, where cost is the number of *add-drop multiplexers* (ADMs). Simulation results are presented comparing the cost of the network with simple lower bounds which show the network has reasonably low cost.

1 Introduction

Optical *wavelength-division-multiplexed* (WDM) networks are expected to become an integral part of telecommunication backbone networks. The networks have fiber optic links, each carrying many communication channels at different wavelengths. The optical channels have high bit rates, e.g., OC-48 (= 2.5 Gb/s) or OC-192 (= 10 Gb/s). At least in the near future, much of telecommunication backbone traffic will be circuit-switched streams being supported by SONET/SDH. These streams will be at a lower-speed (e.g., OC-3 = 155 Mb/s) and must be multiplexed/demultiplexed (i.e., *traffic groomed*) onto and off-of the higher-speed optical channels. It is critical that the network be designed to support expected traffic while keeping network costs at a minimum. In this paper, we will present a WDM ring network architecture that supports *incremental* lower-speed traffic streams, which means that the streams can arrive at arbitrary times but never terminate. (This models the situation when traffic streams are expected to have long holding times, as is usually the case with provisioning high-speed connections today.) The network architecture can be configured to have reasonably low cost.

The network is a WDM ring network (see Figure 1). The ring topology is viable because SONET/SDH self-healing architectures are ring oriented. The links are bidirectional, each composed of two fibers going in opposite directions. The number of nodes

---

*This work was done in part while being with the IBM T.J. Watson Research Center. It was sponsored in part by grant MDA-972-95-C-0001 by DARPA, and was sponsored in part by grant F30602-97-1-0342 by DARPA and Rome Laboratory, Air Force Materiel Command, USAF. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes nothwithstanding any copyright annotation thereon. The views and conclusions contained therein are those of the authors and should not be expressed or implied, of the Defense Advanced Research Projects Agency (DARPA), Rome Laboratory, or the U.S. Government.

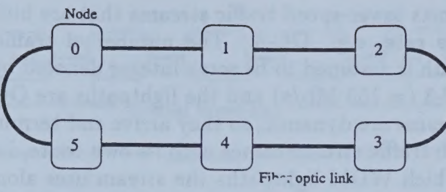


Figure 1: An optical WDM ring.

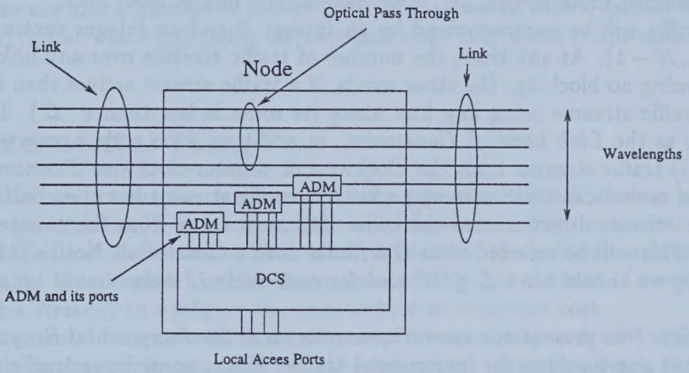


Figure 2: An optical node.

is denoted by N and the nodes are numbered $0, 1, ..., N-1$ in the clockwise direction. The number of wavelengths is denoted by W, and the set of wavelengths is denoted by $\{\omega_0, \omega_1, ..., \omega_{W-1}\}$. On each link, each wavelength ω_i carries a bidirectional (or full-duplex) WDM channel. Notice that a full-duplex WDM channel at a wavelength, say ω_i, on a link is composed of two unidirectional channels at ω_i being carried by the pair of fibers of the link.

A network node is shown in Figure 2. It is composed of *add/drop multiplexers* (ADMs) and a single *digital crossconnect system* DCS. The DCS will be assumed to be *wide sense nonblocking*, which means that a traffic stream may be set-up or terminated through it without disturbing existing traffic streams, assuming availability of access ports. The main contribution of cost at a node will be due to the ADMs and DCS.

Notice that optical signals of a WDM channel may be either terminated by an ADM or *passed through* the node to the next link. If terminated by an ADM, the signal is translated into a high-speed electronic bit stream composed of its lower-speed (tributary) traffic streams. The lower-speed streams may be added and dropped at the node via the ADM interface ports. The interface ports are connected to the main DCS of the node, where the streams may be crossconnected to other ADMs or local access ports. However, if the optical signals of a WDM channel are passed-through the node, then they continue on the next link at the same wavelength (i.e., no *wavelength conversion*). Notice that the traffic carried by the passed-through optical signals are not intended for the node. Also notice that the network supports optical end-to-end communication connections that are terminated by ADMs and are passed-through intermediate nodes. These full-duplex optical connections will be referred to as *lightpaths.* All lightpaths will have a fixed high-speed bit rate (e.g., OC-48 or OC-192).

The network supports lower-speed traffic streams that are bidirectional and, for simplicity, have the same rate, e.g., OC-3. The number of traffic streams that can be supported by a lightpath is assumed to be some integer denoted by c. For example, if the traffic streams are OC-3 (= 155 Mb/s) and the lightpaths are OC-48 (= 2.5 Gb/s) then $c = 16$. The traffic streams are dynamic, so they arrive and terminate at arbitrary times. It is assumed that each traffic stream comes with its own route, i.e., it is *pre-routed.* The network determines which WDM lightpaths the stream uses along its route. Thus, the network has a *lightpath assignment algorithm* (LAA). If a lightpath assignment cannot be found for an arriving traffic stream then it is blocked. The routes for the traffic streams are also assumed to be *simple*, i.e., they traverse any link at most once.

The traffic will be parameterized by an integer $\mathcal{L}$ and an integer vector $t = (t(i) : i = 0, 1, ..., N-1)$. At any time, the number of traffic streams over any link is at most $c \cdot \mathcal{L}$, assuming no blocking. (In other words, if a traffic stream arrives then the number of other traffic streams using any link along its route is less than $c \cdot \mathcal{L}$.) This will be referred to as the *Link Load $\mathcal{L}$ Constraint.* In addition, each node i may terminate at most $c \cdot t(i)$ traffic streams from the clockwise or counter-clockwise direction along the ring. Thus, node i can terminate up to $2c \cdot t(i)$ traffic streams but then half must come from the clockwise direction and the other half must come from the counter-clockwise direction. This will be referred to as the *Nodal Load t Constraint.* Notice that to insure no blocking we should have $\mathcal{L} \leq W$ and for each node i, there should be at least $t(i)$ ADMs.

In Section 2 we present our network, we refer to as the *Incremental Ring Network.* It is wide-sense non-blocking for incremental traffic. Thus, an arriving traffic stream will be able to find a lightpath assignment without disturbing existing traffic streams (under the Link Load and Nodal Load Constraints). We also present a design algorithm that configures the network to minimize cost. The configuration involves placing ADMs at nodes at different wavelengths which, in turn, determines the lightpaths of the network.

The network cost will be the sum of costs at each node. We will assume that the cost of each node i is dominated by the ADM and DCS costs. We will assume that the ADM cost is proportional to the number of ADMs at node i, which we denote by $A(i)$. We will assume that the DCS cost is a function of $A(i) + t(i)$ because $c(A(i) + t(i))$ is an upper bound on the number of (lower-speed) ports at the DCS. For example, the DCS cost could be a quadratic function of $A(i) + t(i)$. Thus, we will assume that the cost at node i is a function $D(A(i), t(i))$, and therefore the cost of the network is $\sum_i D(A(i), t(i))$.

Simulation results are presented in Section 3. The results show the performance of the design algorithm for randomly generated values of t. In Section 4 we state our conclusions.

For the remainder of this section, we will discuss related work. First, we should mention that this paper is a continuation of [8], where a number of WDM ring networks were considered and electronic (ADM) costs were compared. The Incremental Ring network was introduced briefly, but without any analysis or design algorithm. In addition, it was constrained to a uniform traffic model ($t(i)$ is the same for all nodes i).

ADM costs were also considered in [6, 10, 14], but where traffic was assumed to be static. In [6], the traffic was a collection of lightpaths to be set up in a WDM ring network. In [10, 14], the traffic was a *uniform* pattern of lower-speed traffic streams. The static traffic pattern of all traffic streams terminating at one *central* node was also considered in [10].

We should finally note that typically researchers have concentrated on numbers of wavelengths, congestion, delay, or probability of blocking. There is previous work on WDM network design for lower-speed traffic streams [2, 4, 5, 12, 13], assuming traffic is

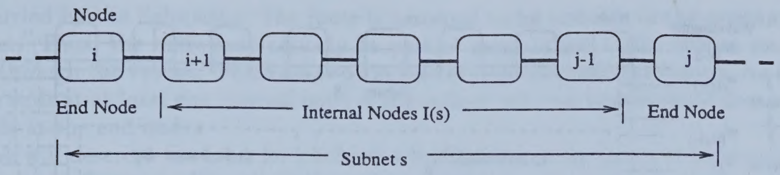


Figure 3: The segment of subnet s.

static. There are also a number of papers on WDM networks with dynamic traffic (e.g., [3, 1, 7, 9, 11]), but where lightpaths are switched and lower speed traffic streams are not considered.

2 Incremental Ring

In this section we present the *Incremental Ring network*, and will assume incremental traffic. The network has $W = \mathcal{L}$ wavelengths which is the minimum to insure no blocking under the Link Load $\mathcal{L}$ Constraint. We will first describe the network architecture, and then show that it is nonblocking for incremental traffic. At the end of the section, we will present a strategy to configure the architecture to minimize cost.

The Incremental Ring network is organized with respect to pieces of it called *subnets.* A subnet s is composed of a sequence of links (and nodes) along the ring in the clockwise direction. This sequence will be referred to as its *segment.* The subnet's *length* is the length of its segment. The subnet's *end nodes* are the nodes at the end of its segment. The rest of its nodes are referred to as the *internal nodes* and denoted by $I(s)$. Figure 3 shows a segment and its end and internal nodes.

A subnet s is also composed of a collection of wavelengths and lightpaths. The wavelengths are $w_0, w_1, ..., w_{r(s)-1}$, where $r(s)$ denotes the number of wavelengths for s. The lightpaths only use the wavelengths and links of s.

The Incremental Ring architecture is organized as a *tree* of subnets. One of the subsets is referred to as the *root* and its segment includes all the links of the ring. The root's segment starts and ends at some node, which we refer to as the *root node.* The number of wavelengths for the root is W, i.e, $r(root) = W$.

If a subnet s has length greater than one (i.e., it has at least one internal node) then it is referred to as a *parent* subnet. Note that the *root* subnet is a parent if $N > 1$. A parent subnet s has two *children* subnets, say s_0 and s_1. The segments of the children are defined by bisecting the segment of s at some internal node denoted by $b(s) \in I(s)$. The resulting subsegments are the segments for s_0 and s_1. The number of wavelengths assigned to each of the children is $r(s_0) = r(s_1) = \rho(s)$, where $\rho(s)$ is defined to be

$$\rho(s) = \min\{r(s), \sum_{i \in I(s)} t(i)\}. \tag{1}$$

Note that $\rho(s)$ is equal to

$$\min\{W, \sum_{i \in I(s)} t(i)\} \tag{2}$$

since $r(root) = W$ and for an arbitrary parent subnet s' and its child s^*, $\sum_{i \in I(s')} t(i) \geq \sum_{i \in I(s^*)} t(i)$. Figure 4 shows an example subnet s, its bisecting node $b(s)$, and two children s_0 and s_1. As shown in the figure, the $r(s)$ wavelengths of s are categorized into two types for s: *internal* and *transit.* The internal wavelengths of s are $\{w_0, w_1, ..., w_{\rho(s)-1}\}$,

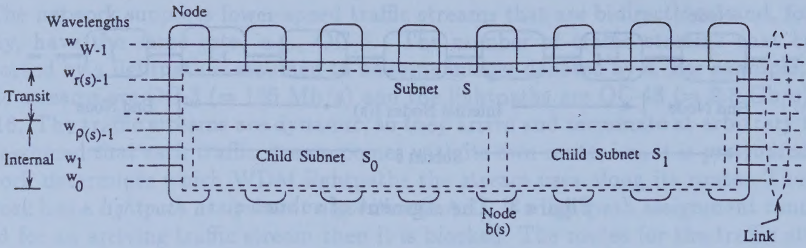


Figure 4: A parent subnet s composed of its two children subnets s_0 and s_1.

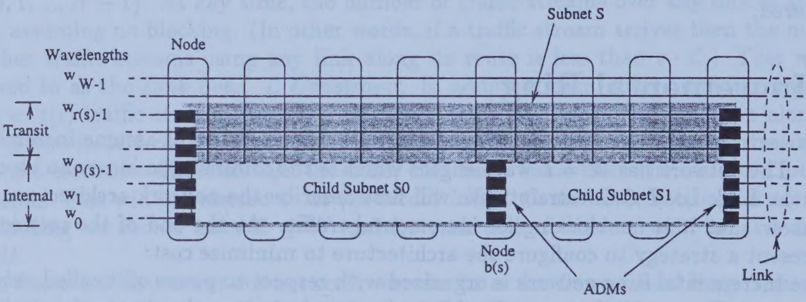


Figure 5: A parent subnet s composed of its two children subnets s_0 and s_1 and the placement of ADMs.

and the transit wavelengths of s are $\{w_{\rho(s)}, w_{\rho(s)+1}, ..., w_{r(s)-1}\}$. Thus, the $\rho(s)$ internal wavelengths of s are the wavelengths of its children. This categorization of wavelengths will be used later to define where lightpaths are placed in the subnet (or equivalently where ADMs are placed), and how lightpaths are assigned to traffic streams.

If a subnet s has only one link then it has no children and is called a *leaf* subnet. Leaf subnets only have transit wavelengths. Now note that the collection of leaf and parent subnets form a tree of subnets, where the root and leaf subnets are the root and leaves, respectively, of the tree. Also note that a subnet can be both a child (of some subnet) and a parent (of two other subnets).

Next, we define how ADMs are placed at each node, which will in turn define the placement of lightpaths. First, the root node has W ADMs terminating each wavelength on both of its links. For all other nodes, the placement is as follows. A node must be a bisecting node $b(s)$ for some parent subnet s. The node $b(s)$ has $\rho(s)$ ADMs, which terminate wavelengths $w_0, ...w_{\rho(s)-1}$ on each of the node's links. Thus, for subnet s, its end and bisecting nodes terminate lightpaths on its internal wavelengths, while *only* its end nodes terminate lightpaths on its transit wavelengths (refer to Figure 5). Thus, lightpaths on its transit wavelengths directly connect its end nodes, and cannot be used for traffic streams that terminate at its internal nodes.

To complete the description of the Incremental Ring architecture, we define its *lightpath assignment algorithm* (LAA). The LAA takes an incoming traffic stream with a route, and finds a contiguous sequence of lightpaths along the route that can carry the stream. To define the LAA for the ring, we will define LAAs for the subnets.

The LAA for a subnet takes a route and assigns a contiguous sequence of its lightpaths along the route that have spare capacity. Thus, a traffic stream following the route can be carried by the lightpaths. The route is assumed to be entirely in the segment of the subnet. Thus, the route may actually be part of some larger traffic stream route that goes through the subnet. There are two kinds of routes: *internal* and *transit*. An internal route ends at at least one internal node of the subnet, while a transit route does not, i.e., it ends at the end nodes.

We will describe the LAA for a subnet s by first assuming that s is a parent subnet, as shown in Figure 4. How the LAA assigns lightpaths to a route depends on whether the route is internal or transit. If the route is internal then it is first split at the bisecting node $b(s)$. (Of course, it is not split if it does not pass through $b(s)$.) The resulting subroutes fit into the segments of the children subnets of s. The LAAs of the children are used to assign lightpaths to the subroutes. These assignments combine to provide the assignment for the original route. However, if the route is transit, then the LAA will first attempt to assign a lightpath on a transit wavelength of s to the route. If all such lightpaths have no spare capacity then the LAA assigns lightpaths to the route as if it were an internal route, i.e., it splits the route at $b(s)$ and uses the LAAs of the children subnets.

The LAA for a leaf subnet s is simpler because all routes and wavelengths are transit. A route is assigned to a lightpath of s with spare capacity.

We have completed the definition of the LAAs for subnets. The LAA for the root subnet is basically the LAA for the Incremental Ring. In particular, the LAA for the Incremental Ring will assign lightpaths to an incoming traffic stream as follows. It first splits the route of the stream at the root node (if the route passes through the root). The resulting subroutes are entirely in the root subnet and can be assigned lightpaths by the subnet's LAA. This lightpath assignment is an assignment for the traffic stream since the lightpaths that meet at the root node can be crossconnected.

Next we will prove the following property.

Theorem 1 *Consider an Incremental Ring network under Traffic Assumption B and incremental traffic. Then the Incremental Ring network is wide sense nonblocking.*

The theorem is implied by the next lemma, which uses the following definitions. Consider a subnet s. The subnet is said to have *spare capacity* in a link if one or more of its wavelengths have spare capacity in the link. The subnet is said to be *wide sense nonblocking* if given a route, the LAA of the subnet will find a lightpath assignment for the route as long as the subnet has spare capacity in each link along the route. Note that having the spare capacity does not preclude a valid lightpath assignment because in order for a set of lightpaths to carry a traffic stream they must terminate at the ends of the route, and at common intermediate nodes. Note that the next lemma implies Theorem 1 because it is applicable to the root subnet.

Lemma 1 *Consider an Incremental Ring network with incremental traffic. Then each subnet s of the ring is wide sense nonblocking.*

Proof. To prove the lemma for leaf subnets is trivial since the subnet has only one link. Thus, we will only consider parent subnets. The proof of the lemma is by induction. Consider a parent subnet s and its two children subnets s_0 and s_1. We will assume that the children subnets are nonblocking, and proceed to show that s is also nonblocking.

Let us see how subnet s assigns lightpaths to traffic streams over time. Initially, there are no traffic streams so all lightpaths in the subnet are empty. As traffic streams arrive they are assigned lightpaths so that internal traffic (i.e., streams that terminate at internal nodes of s) are assigned to lightpaths with internal wavelengths, and transit traffic (i.e., streams that do not terminate at internal nodes of s) are assigned to lightpaths with transit wavelengths.

Note that this can continue while transit traffic does not completely fill lightpaths with transit wavelengths. This is partly due to the traffic model which assumes that each internal node $i \in \mathcal{I}(s)$ can terminate at most $c \cdot t(i)$ traffic streams from the clockwise or counter-clockwise direction. This implies that for any link of the subnet, the number of internal traffic streams that go over it is at most $c \cdot \rho(s)$. The internal traffic can be accommodated by the lightpaths with internal wavelengths because

- children subnets are assumed to be nonblocking and occupy the $\rho(s)$ internal wavelengths; and
- bisecting node $b(s)$ (which is between the children subnets) terminates all the internal wavelengths with ADMs, and can therefore crossconnect the spare capacity in the two children subnets.

Thus, while transit streams do not completely fill the transit wavelengths, internal traffic streams will be accommodated by lightpaths at internal wavelengths. Hence, there will be no blocking.

Now at some point the lightpaths with transit wavelengths may become completely filled. Then the only spare capacity left in the subnet s is in its internal wavelengths. Arriving traffic streams will not be blocked from this spare capacity because (i) the spare capacity resides in the nonblocking children subnets, and (ii) the bisecting node $b(s)$ can crossconnect the spare capacity in the two children subnets. □

Finally, we will describe a strategy to place ADMs (and lightpaths) to minimize cost. Notice that the placement of ADMs is dependent on the choice of root and bisecting nodes. Thus, the strategy is to find an optimal choice of these nodes. Before presenting the strategy, we give some definitions.

- Let $\sigma(i,k)$ denote a segment of the ring that starts at node i and goes clockwise over k links.
- Let $I'(i,k)$ denote the *internal* nodes of $\sigma(i,k)$, i.e., $I'(i,k) = \{(i+j) \bmod N : 0 < j < k\}$.
- Consider all possible ways of configuring the Incremental Ring, i.e., all possible ways of choosing the root and bisecting nodes. Of these, consider all configurations that have a subnet with segment $\sigma(i,k)$. Let $C(i,k)$ denote these configurations of the Incremental Ring.
- For each configuration $\gamma \in C(i,j)$, let $q_\gamma(i,k)$ be the sum of costs over the internal nodes $I'(i,k)$ assuming the Incremental Ring is configured as γ. Let $q(i,k) = \min_{\gamma \in C(i,j)} q_\gamma(i,k)$. Thus, $q(i,k)$ is the minimum sum of costs at the internal nodes $I'(i,k)$ assuming that $\sigma(i,k)$ is the segment of a subnet.

Notice that for each node i, $D(W,t(i)) + q(i,N)$ is the minimum cost for the Incremental Ring assuming i is the root node because the root has cost $D(W,t(i))$ and the minimum cost of the other nodes is $q(i,N)$. Thus, the minimum cost for the Incremental Ring (over all possible roots) is $\min_i[D(W,t(i)) + q(i,N)]$.

We now describe the strategy to determine the root and bisecting nodes. It consists of two steps. The first step is to compute the values for $q(i,k)$ for all nodes i and all $k = 1,2,...,N$ by using the equation

$$q(i,k) = \begin{cases} 0, & \text{if } k = 1 \\ \min_{i<j<k} \{q(i,j) + q((i+j) \bmod N, k-j) + d(i,j)\} & \text{if } k > 1 \end{cases} \quad (3)$$

where

$$d(i,j) = D\left(\min\{W, \sum_{n\in I'(i,k)} t(n)\}, \quad t((i+j) \bmod N)\right) \tag{4}$$

The Equation (3) is true for $k = 1$ because $\sigma(i,1)$ corresponds to a leaf subnet, which has no internal nodes (and no ADMs or DCS). The equation is also true for $k > 1$ because (a) the minimum is over all bisecting nodes $(i+j) \bmod N$ for a subnet with segment $\sigma(i,k)$, (b) the terms $q(i,j)$ and $q(i+j,k-j)$ are the minimum cost at internal nodes of the two children of the subnet, and (c) the term $D\left(\min\{W, \sum_{n\in I'(i,k)} t(n)\}, t((i+j) \bmod N)\right)$ is the cost at the bisecting node where $\min\{W, \sum_{n\in I'(i,k)} t(n)\}$ is the number of its ADMs. The values for $\{q(\cdot,\cdot)\}$ are determined by first computing $q(i,2)$ for all nodes i. Then computing $q(i,3)$ for all nodes i, and so forth.

The second step is to determine the root and bisecting nodes that minimize the cost of the network from the values of $\{q(\cdot,\cdot)\}$. This is straightforward. In particular, the node i that minimizes $\min_i q(i,N)$ is the optimal root node. The node $(i+j) \bmod N$ that minimizes the right hand side of (3) is the optimal bisecting node for the corresponding subnet.

3 Simulation Results

In the previous section, a design algorithm was given to determine the placement of ADMs for the Incremental Ring Network. In this subsection, simulation results of the algorithm will be given. For simplicity, the cost of a node will be the number of its ADMs. Thus, DCS costs are ignored. We will compare the cost resulting from the design algorithm with the simple lower bound $\frac{1}{N}\sum_{i=0}^{N-1} t(i)$. For the simulations, we assume that $W = 32$ and $c = 16$ (e.g., 32 OC-48 WDM channels per link, and traffic streams are OC-3).

Each simulation run corresponds to a randomly generated set of values for $(t(i) : i = 0,1,...,N-1)$. This is determined by a single parameter $F \geq 1$. To generate the values, first a collection $f = (f_i : i = 0,1,...,N-1)$ of real values are randomly generated. The values are independent and uniformly distributed over the interval $[1,F]$. The value f_i represents the relative frequency that node i will terminate a traffic stream. For example, if $f_i/f_j = 2$ then node i will be twice as likely to terminate a traffic stream as node j. Next, given f, a collection of traffic streams is generated. The collection will fill each link with exactly $c \cdot W$ streams, thereby maximally loading each link. The streams are generated sequentially and go in the clockwise direction around the ring. The first stream starts at a random node. It is routed clockwise and its end node will be at most $N/2$ hops away (to model shortest path routing of streams). Its end node is chosen randomly according to f. Thus, node i will be f_i/f_j more likely to be chosen than node j. There is the exception that the node that is exactly $N/2$ hops away has its frequency decreased by a half, to take into account that a shortest hop path between two nodes at opposite ends of the ring could be on either side of the ring. Now the end node of the first traffic stream becomes the starting node of the second traffic stream. The end node of the second traffic stream is randomly chosen using f. The end node of the second traffic stream becomes the starting node of the third traffic stream, and so on. This continues until all bandwidth in each link is accounted for. After the collection of traffic streams has been generated, each $t(i)$ is computed. In particular, if $g(i)$ is the number of traffic streams terminating at node i from the clockwise (or counter clockwise) direction on the ring, then $t(i) = \lceil g(i)/c \rceil$.

We ran simulations for $N = 8, 12$, and 16 for different values of F. Figure 6 shows the average values of

- $ADMAVG =$ the average number of ADMs per node,
- $tAVG =$ the average value of $t(i)$ (i.e., $\frac{1}{N}\sum_{i=0}^{N-1} t(i)$), which is a lower bound on the average number of ADMs per node.
- $tMIN =$ the minimum value of $t(i)$ (i.e., $\min_i t(i)$),
- $tMAX =$ the maximum value of $t(i)$ (i.e., $\max_i t(i)$).

as functions of F for $N = 16$, where each data point is an average over 100 random instances.

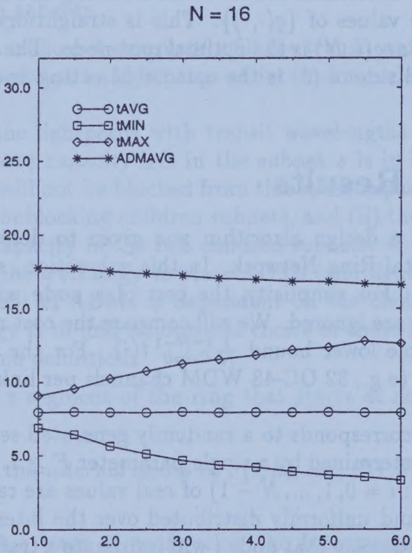


Figure 6: For a ring network with $N = 16$ and $W = 32$, the plots give the average values of $tAVG$, $tMIN$, $tMAX$, and $ADMAVG$

Note that as F increases the difference in value between $tMIN$ and $tMAX$ increases, as expected. In addition, the number of ADMs required decreases slightly.

Figures 7 and 8 show the ratios $\frac{ADMAVG}{tAVG}$ and $\frac{ADMAVG}{W}$, respectively, for $N = 8, 12$, and 16. The first ratio indicates the cost (in ADMs) of the Incremental Ring relative to the simple lower bound. Note that it is approximately two, and thus traffic streams will on average traverse approximately two lightpaths. This is reasonably low especially if traffic streams must be groomed to utilize bandwidth. The other ratio $\frac{ADMAVG}{W}$ indicates a cost savings due to optical signals being allowed to pass through nodes (i.e., not terminating every wavelength). Note that the cost savings is at least 27%, 36%, and 44% for $N = 8, 12$, and 16, respectively (see Figure 8). This can be considerable (e.g., if ADMs cost \$100,000 per node then a 27% savings amounts to saving \$864,000 per node, and a 44% savings amounts to saving \$1.4 million per node). Both Figures 7 and 8 show that the cost of the Incremental Ring improves as F increases.

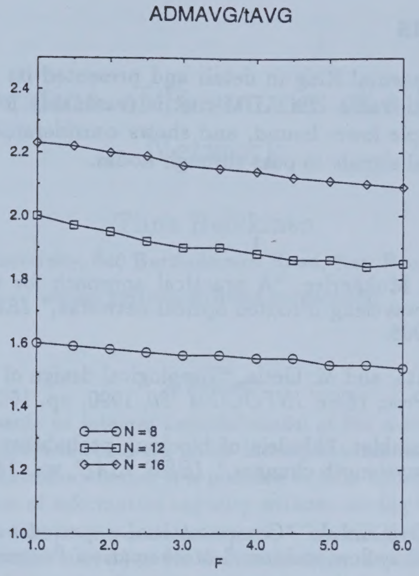


Figure 7: The ratio $ADMAVG/tAVG$ as a function of F for three different values of N. The curves from top to bottom are for $N = 16, 12$, and 8, respectively.

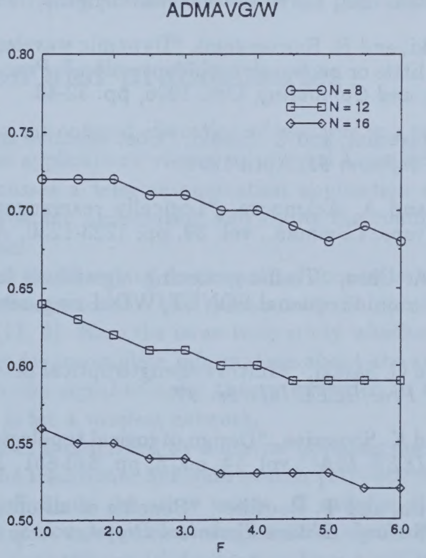


Figure 8: The ratio $ADMAVG/W$ as a function of F for three different values of N. The curves from top to bottom are for $N = 8, 12$ and 16, respectively.

4 Conclusions

We described the Incremental Ring in detail and presented its wide sense nonblocking property for incremental traffic. Its ADM cost is reasonably low, approximately twice the value of a very simple lower bound, and shows considerable savings over networks that do not allow optical signals to pass through nodes.

References

[1] D. Banerjee and B. Mukherjee, "A practical approach for routing and wavelength assignment in large wavelength-routed optical networks," *IEEE JSAC*, vol. 14, no. 5, June 1996, pp. 903-908.

[2] J. Bannister, L. Fratta, and M. Gerla, "Topological design of the wavelength-division optical networks," *Proc. IEEE INFOCOM '90*, 1990, pp. 1005-1013.

[3] R. Barry and P. Humblet, "Models of blocking probability in all-optical networks with and without wavelength changes," *IEEE JSAC*, vol. 14, no. 5, June 1996, pp. 858-865.

[4] D. Bienstock and O. Gunluk, "Computational experience with a difficult mixed-integer multicommodity flow problem," *Mathematical Programming,* vol. 68, pp. 213-237, 1995.

[5] A. Ganz and X. Wang, "Efficient algorithm for virtual topology design in multihop lightwave networks," *IEEE/ACM Trans. Networking*, vol. 2, pp. 217-225, June 1994.

[6] O. Gerstel, P. Lin, and G. Sasaki, "A new angle on wavelength assignment in WDM rings: minimize system cost, not number of wavelengths," *Proc. IEEE Infocom 98.*

[7] O. Gerstel, G. Sasaki, and R. Ramaswami, "Dynamic wavelength allocation in WDM ring networks with little or no wavelength conversion," *Proc. 1996 Allerton Conf. on Commun., Control, and Computing*, Oct. 1996, pp. 32-43.

[8] O. Gerstel, R. Ramaswami, and G. Sasaki, "Cost effective traffic grooming in WDM rings," *Proc. IEEE Infocom 98*, April 1998.

[9] J.-F. Labourdette and A. Acampora, "Logically rearrangeable multihop lightwave networks," *IEEE Trans. Commun.,* vol. 39, pp. 1223-1230, Aug. 1991.

[10] E. Modiano and A. Chiu, "Traffic grooming algorithms for minimizing electronic multiplexing costs in unidirectional SONET/WDM ring networks," *Proc. CISS '98, to appear.*

[11] R. Ramaswami and G. Sasaki, "Multiwavelength optical networks with limited wavelength conversion," *Proc. IEEE Infocom 97.*

[12] R. Ramaswami and K. Sivarajan, "Design of logical topologies for wavelength-routed optical networks," *IEEE JSAC*, vol. 14, no. 5, pp. 840-851, June 1996.

[13] P. Roorda, C.-Y. Lu, and T. Boutilier, "Benefits of all-optical routing in transport networks," *OFC '95 Conf. Edition Technical Digest*, vol. 8, pp. 164-165, 1995.

[14] J. Simmons, E. Golstein, and A. Saleh, "On the value of wavelength-add/drop in WDM rings with uniform traffic," *Proc. OFC '98*, pp. 361-362, Feb. 1998.

On the Allocation of Resources of a Multi-Media Network

Tiina Heikkinen
RUTCOR, Rutgers University, 640 Bartholomew Road, New Brunswick, NJ 08854-8003
email heikkine@rutcor.rutgers.edu

Abstract

This paper presents an integral Leontief model of the max-min fair allocation of resources of a multimedia network under incomplete information. A motivation for the study is to examine whether it is possible to infer about the optimality of a particular realization of information capacity without solving a completely known system describing a telecommunication network. This will give conditions for solving a multimedia (multireceiver) power control problem as a discrete optimization problem, i.e. by each datatype/receiver separately such that this leads to the same solution as a centralized resource allocation in the whole network. The key result turning the resource allocation problem into a discrete optimization problem is the general separation, due to Leontief structure, between optimal power allocation and channel allocation under uncertainty. The greedy algorithm is the optimal algorithm for the centralized problem, defined over a polymatroidal structure.

1 Introduction and preliminaries

The topic of this paper is the optimal allocation of resouces in a multimedia network with Leontief structure. Some applications viewed as integral Leontief problems are discussed in [7]. This paper discusses a telecommunication application modelled using a more general Leontief-von Neumann formulation and shows the conditions for integrality for such a generalized system.

Standard results from the theory of mathematical economics show that the optimization of the allocation of resources in the system may be decomposed into subsidiary optimization problems [11, 6]. Here the issue is to study whether the decomposition result continues to hold under incomplete information about the state of the system. The framework of analysis is the signal-to-noise theory extended to the case of multimedia. The system description is for a wireless network.

The state of the system is described by a matrix of fading coefficients, **G**, where **G** is an enlarged version of the traditional link gain matrix [17] such that each user is allowed to have several gain coefficients. In other words, **G** satisfies the condition of being a Leontief matrix with at most one positive component in each column of datatypes, over all rows of users. ([5] derives this model description from a utility theoretical setting.)

---

[1] This work was supported by ONR under contract number N000014-92-3-1375 and in part by the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) which was funded by the NSF under contract STC-91-19999

Denote the vector of transmit powers by $\mathbf{x} = (x_1, ..., x_{nm})'$, with n users and m datatypes. The received signal vector can thus be described by $\mathbf{Gx}$. The signal-to-noise theory makes the assumption is that the received signals of other users are noise, called an externality cost and described $\mathbf{Fx}$ where $\mathbf{F}$ denotes the matrix of the interfererence coefficients.

It will be shown that a centralized solution under incomplete information and exernalities is necessary to find the globally optimal transmit power allocation; a decentralized solution corresponds to the centralized solution under complete information. If there is no information advantage in the centralized formulation of the problem as compared to the decentralized formulation, then the decentralized and centralized solutions are the same. A greedy (individually optimal) algorithm is optimal to solve the centralized problem defined on a polymatroid.

Another interpretation of the model is to consider the various datatypes as representing multiple receivers or transmitters. The implication of the model is that it is possible to solve the individual subproblems in various receivers separately; a centralized solution is not necessarily needed. A key result is a separation between optimal channel (datatype) allocation and power allocation; irrespective of the channel allocation, the optimal power control rule remains the same.

The method of analysis under incomplete information is the theory of simple games [14]. Under incomplete information about the capacity game inference about the optimality is based on a set of observations. Based on the observations it is possible to generate a class of exchangeable games with the same probability distribution as with the observed game (section 3.1). Section 3.2 studies the case where games are not exchangeable.

Sections 2 – 3 develop a theory of integer-valued user capacity. Section 2 presents a centralized solution under complete information (2.1) and conditions under which a decentralized solution coincides with the centralized solution (2.2). Section 3 seeks to study optimal capacity of a multimedia network under incomplete information. Section 4 studies the optimality of the greedy algorithm and 5 presents a conclusion.

2 A centralized solution to the allocation of resources in a multi-media network

In [5] the capacity problem of finding optimal transmit powers for a multimedia network is formalized using the von Neumann growth model Here due to the block-symmetric/diagonal structure of the gain matrix $\mathbf{G}$, (and the corresponding structure in the interference matrix $\mathbf{F}$), an integral Leontief-von Neumann model applies.

2.1 A centralized solution based on nonsubstitution theorem

The nonsubstitution theorem states that the optimal capacity can be found by solving a subproblem of the original von Neumann model:

$$\max L_{\mathbf{x}} = \max_{\mathbf{x}} \min_{\mathbf{p}} (\alpha - \mathbf{p}'(\alpha \mathbf{Fx} - \mathbf{Gx})), \tag{1}$$

where α denotes the information capacity. The subproblem corresponds to finding the subsystem with the least dominant characteristic root of the modified interference matrix.

Theorem 1 *((Non)substitution theorem) The von Neumann model has a unique optimal capacity α^*, and unique power and state vectors $\mathbf{x}^*$, $\mathbf{p}^*$. Here $\alpha^* = \frac{1}{\lambda}$, where λ is the least dominant characteristic root of a square submatrix of $\mathbf{G}^{-1}\mathbf{F}$.*

Proof. Straightforward generalization of [8].

Intuitively, what the nonsubstitution theorem says is that even if it was possible to consider the multimediatypes as substitutable, in an extreme case of linear utility ($\mathbf{Gx}$), substitution is not optimal (see e.g. [11]). (This result is valid for more general utility functions u of user i from signals, see [5]).

Proposition 1 *The optimal capacity α^* in a multimedia network is the same for all users and is given by*

$$\alpha^* = \frac{1}{\lambda} = \min_{x \in M^n} \max_i \frac{g_i x_i}{(\mathbf{Fx})_i}. \tag{2}$$

Proof. First define the least dominant characteristic root λ by

$$\lambda = \max_{x \in M^n} \min_i \frac{(\mathbf{Fx})_i}{g_i x_i}, \tag{3}$$

where $M^n = \{x | x \geq 0, \sum x_i = c\}$. Let $(\mathbf{Gx})_i$ denote the ith component of $\mathbf{Gx}$. We need to show that

$$\sup_i \frac{(\mathbf{Gx})_i}{(\mathbf{Fx})_i} = \frac{1}{\inf_i \frac{(\mathbf{Fx})_i}{(\mathbf{Gx})_i}}. \tag{4}$$

Let

$$\inf_i \frac{(\mathbf{Fx})_i}{(\mathbf{Gx})_i} = \frac{1}{\alpha}. \tag{5}$$

Thus,

$$\forall \epsilon > 0, \frac{(\mathbf{Fx})_i}{(\mathbf{Gx})_i} \geq (\frac{1}{\alpha} - \epsilon)(\mathbf{Gx})_i. \tag{6}$$

By the definitions of $\mathbf{F}, \mathbf{G}$ and M^m, $\alpha \geq 0$. Further $\inf \frac{(\mathbf{Fx})_i}{(\mathbf{Gx})_i}$ is finite. Let $0 < \frac{1}{\beta} < \frac{1}{\alpha}$. Then $(\mathbf{Fx})_i \geq \frac{1}{\beta}(\mathbf{Gx})_i$. That is, $(\mathbf{Gx})_i \leq \beta(\mathbf{Fx})_i$ and $(\mathbf{Gx})_i \geq \beta(\mathbf{Fx})_i$ if $\beta \leq \alpha$. This is equivalent to: $\sup \frac{(\mathbf{Gx})_i}{(\mathbf{Fx})_i} = \alpha$. Finally, suppose $\frac{1}{\alpha} = 0$. Then $(\mathbf{Fx})_i \geq \frac{1}{\beta}(\mathbf{Gx})_i$ fails to hold $\forall \frac{1}{\beta} > 0$. Thus $(\mathbf{Gx})_i \leq \beta(\mathbf{Fx})_i$ fails to hold $\forall \frac{1}{\beta} > 0$. This implies that $\sup \frac{(\mathbf{Gx})_i}{(\mathbf{Fx})_i} = \infty$.

Similarly,

$$\inf_{\mathbf{x}} \frac{(\mathbf{Gx})_i}{(\mathbf{Fx})_i} = \frac{1}{\sup_{\mathbf{x}} \frac{(\mathbf{p}, \mathbf{Gx})_i}{(\mathbf{p}, \mathbf{Fx})_i}}. \tag{7}$$

Since in a square model $(\mathbf{Gx})_i = g_i x_i$, the result follows from *theorem 1*, extending the optimum power control rule in [17] to the case of a multimedia (multireceiver) network.

Also notice that here, due to the relation between $\mathbf{G}$ and $\mathbf{F}$, the following holds: *Proposition 2* allows to replace the distribution of α by the distribution of the inverse of the number of users (in section III.C).

Proposition 2 *Assume $\mathbf{G}$ is a Leontief matrix with at most one positive component in each column. Then λ evaluated at optimal x^* is integral.*

Proof. By the assumption of the proposition, optimal capacity is one corresponding to a square system with equal optimal capacities. Since the optimal capacity α^* is the same for users, the received signal $X_i = g_i x_i^* = X$ is the same for users: $\frac{1}{\alpha}\mathbf{X} = \mathbf{AX}$, where $\mathbf{A}$ is matrix consisting of zeros on diagonal and ones elsewhere; hence $\mathbf{A}$ is totally unimodular with maximal eigenvalue equal to rank(A)-1. Thus, $\frac{1}{\alpha}$ is an integer $n-1$. Thus,

$$\lambda^* = \frac{(n-1)X}{X} = n-1. \tag{8}$$

Proposition 3 *Assume that the inverse of the dual of capacity maximization problem has an integral optimum solution. Then this is the optimal user capacity* λ^*, *i.e. the optimum user capacity is given by*

$$\lambda^* = \frac{1}{\alpha^*} + 1. \tag{9}$$

Proof. The dual to capacity optimization is to min $\alpha s.t. \frac{1}{\alpha}\mathrm{p}'\mathrm{G} - \mathrm{p}'\mathrm{F} \leq 0$). The inverse to this problem is given by

$$\min \frac{1}{\alpha} \quad s.t. \quad \frac{1}{\alpha}\mathrm{p}'\mathrm{G} - \mathrm{p}'\mathrm{F} \geq 0. \tag{10}$$

Assume this problem has an integer solution. Then by *Theorem 1* the optimum is found from a square subsystem:

$$\min \frac{1}{\alpha} \quad s.t. \frac{1}{\alpha} = \mathbf{p}'\mathbf{A}, \tag{11}$$

where the ijth component of $\hat{\mathbf{A}}$ is given by $\frac{G_{ij}}{G_{ii}}$ (by *Proposition 2* set $G_{ij} = G_{ii} = \max\{G_{ij}\}_{j \neq i}$). By total dual integrality (TDI) [15], an integral solution to problem (12) implies an integral solution to its dual, the user capacity maximization problem (and vice versa), given below:

$$\max \frac{1}{\alpha} \quad s.t. \quad \frac{1}{\alpha}\mathbf{GX} - \mathbf{FX} \leq 0, \tag{12}$$

which can be rewritten as

$$\max \frac{1}{\alpha} \quad \mathbf{AX} \leq \frac{1}{\alpha}. \tag{13}$$

The system $\mathbf{AX} = \frac{1}{\alpha}$ is totally dual integral if the minimum in

$$\max\{\frac{1}{\alpha}|\mathbf{AX} \leq \frac{1}{\alpha}\} = \min\{\frac{1}{\alpha}|\mathbf{p} \geq 0; \frac{1}{\alpha} = \mathbf{p}'\mathbf{A}\} \tag{14}$$

has an integral optimum solution $\mathbf{p}$ for each integral vector $\frac{1}{\alpha}$ for which the minimum is finite, and also then the maximum is integral.

2.2 A decentralized solution to power control

Another expression for the optimal capacity α^* as derived in [5] is given by

$$\alpha^{**} = \min_{\mathbf{p}} \max_{\mathbf{x}} \frac{\mathbf{p}'\mathbf{Gx}}{\mathbf{p}'\mathbf{Fx}} = \max_{\mathbf{x}} \min_{\mathbf{p}} \frac{\mathbf{p}'\mathbf{Gx}}{\mathbf{p}'\mathbf{Fx}}. \tag{15}$$

Notice that the expression for α^{**} is homogeneous of degree zero [11]; for any $h \geq 0$

$$\alpha^{**}(h\mathrm{p}, h\mathrm{x}) = \alpha^{**}. \tag{16}$$

Thus, it is possible to normalize the vector p so that $\sum p_i = 1$ without effect on the optimum. This allows to consider p as a probability distribution in what follows.

Suppose nature chooses the link gain elements and a probability distribution p on various user types, where type i refers to the spatial distribution of the user i as designated by $(\mathbf{Gx})_i$.

Proposition 4 *Suppose nature chose a game with $\underline{n}$ users, and that the game is common knowledge (the payoffs, players are common knowledge). Then*

$$\alpha^* = \alpha^{**}. \tag{17}$$

Proof. By [15] α^{**} solves the centralized optimum information capacity problem defined above.

The difficulty with the decentralized solution is that users must know the game against nature, i.e. that it is equally likely that nature chooses user i to be of any type j in order for α^{**} to emerge as the noncooperative optimum solution (c.f. stochastic utility maximization [1]). An extreme case where the centralized and decentralized solutions would not coincide would be if the utility functions are not simple linear mappings in capacities [5]; then in order for a decentralized solution to be optimal it would in addition be required that each user knows the utility functions of the other users.

However, the decentralized formalism implies that α^{**} by definition satisfies the criterium of *max-min-fairness* which is the most commonly used fairness criterium in telecommunication networks.

3 Inference of max-min optimality of capacity under incomplete information

This section seeks to define a centralized model of max-min fair resource allocation under incomplete information. It is the analysis of fairness, in particular, that makes *TDI*-property useful. Finding an integral user capacity implies the existence of an extended minimax game defined over all the datatypes:

Proposition 5 *Corresponding to each integral solution to user capacity optimization there exists a set of enlarged matrix games in terms of which the solution is max-min optimal.*

Proof. Follows from *Proposition 3* and equation (15).

Corollary 1 *Any optimum capacity as power control solution to a square subsystem equals to a max-min-fair Pareto-optimal capacity of a class of enlarged systems with several datatypes (receivers).*

3.1 Partial inference in a static model of capacity allocation

Suppose now that there is incomplete information about the max-min game against nature; in other words, assume that the matrices $\hat{\mathbf{G}}$,$\hat{\mathbf{F}}$ represent only one particular realization of the spatial distribution of users; $\mathbf{G}, \mathbf{F}$ are matrix valued random processes. Then the corresponding $\hat{\alpha^*}$ is max-min fair only for a particular realization. However, by *Corollary 1*, the user capacity $\frac{1}{\hat{\alpha^*}}$ corresponds to the capacity defined for a class of more general games. Consider again the problem of nature to choose a probability distribution $\mathbf{p}$ so as to minimize the user capacity:

$$\min \frac{1}{\alpha} \quad s.t. \quad \frac{1}{\alpha}\mathbf{p}'\hat{\mathbf{G}} - \mathbf{p}'\hat{\mathbf{F}} \geq 0. \tag{18}$$

Thus, if nature's problem has an integral solution, the user capacity maximization (the dual) has an integral optimum. Whether the integral solution exists, depends on whether nature chooses $\mathbf{G}$ to satisfy the conditions of being a Leontief-matrix to make the substitution theorem apply.

Proposition 6 *The minimax probability distribution nature chooses over the rows of Leontief matrix* $\hat{\mathbf{G}}, (\hat{\mathbf{F}})$ *is the same as the distribution chosen for the rows of* $\hat{\mathbf{G}} + \mathbf{H}$, *where* $\mathbf{H}$ *is a nonnegative matrix such that* $(\hat{\mathbf{G}} + \mathbf{H})_i \leq \max_i(\hat{\mathbf{G}})_i$.

Proof. Follows from the substitution theorem.

Definition 1 *We say that the random variables* $Y_1, Y_2, ..., Y_n$ *are exchangable when the probability distribution of the random vector* $Y_{i1}, Y_{i2}, ..., Y_{in}$ *is identical to that of* $Y_1, Y_2, ..., Y_n$ *for all permutations* $(i_1, i_2, ..., i_n)$ *of the subscripts* $(1, 2, ..., n)$. *Furthermore, an infinite sequence* Y_1, Y_2 *of random variables is said to be exchangable if each finite segment* $Y_1, Y_2, ..., Y_n$ *constitutes exchangable random variables [14].*

Proposition 7 *A subset of exchangeable binary random variables is also exchangeable.*

Proof. See [14].

Definition 2 *Define* $A = \{(\mathbf{G} + \mathbf{H}_i)\}_i^K$ *to be the set of matrix games generated by* $\hat{\mathbf{G}}$ *if* $\mathbf{H}_i$ *is a nonnegative matrix such that* $(\hat{\mathbf{G}} + \mathbf{H}_i)_j \leq \max_j(\hat{\mathbf{G}})_j$. *Let* $\hat{\alpha}_i$ *denote the capacity of matrix game* $A_i \in A$, *and let* $\theta_i = 1$ *if* $\hat{\alpha}_i \geq \bar{\alpha}$ *and zero otherwise.*

Thus, a sequence of random variables $(\theta_i)_j^J$ is exchangable. In conclusion, every integral capacity is max-min fair relative to a class of generated games. Even if one chooses to solve power control using a square model with unkown elements of the gain and interference matrices, the solution will be max-min optimal for some class of matrix games.

3.2 The derivation of a reliability function

To begin with, define a partially defined user capacity:

Definition 3 *A partially defined user capacity is a boolean mapping* $f : 2^k \rightarrow \{0, 1\}$, *where* k *denotes the number of users in the system and* $f = 1$ *if the system is reliable, i.e, more than* p *percent of users reach the optimal capacity and zero otherwise.*

The issue in both defining the static and the temporally optimal capacity is in extending the partially defined capacity to the case of the whole sample space. In other words, to be able to define $\bar{\alpha}$ it is necessary to define *completely defined user capacity*

Definition 4 *A completely defined user capacity is a mapping* $F : 2^N \rightarrow \{0, 1\}$, *where* N *is the upper bound for the number of users and* f *is one when all users find the optimal capacity and 0 otherwise.*

A way of finding a completely defined user capacity is to extend the theory of simple games [14].

Definition 5 *Define a winning coalition to be the set of users accomodated in the system.*

Consider a proper simple game where each winning coalition is a blocking coalition. Winning coalitions form when nature chooses the gain coefficients positive for a set of users for which $\hat{\alpha^*} \geq \bar{\alpha}$ or $\alpha_t \geq \bar{\alpha}$. Obviously there cannot be another winning coalition in the temporal model (that period, the other users lost).

Let N be a finite set of users (players) in the area covered.

Definition 6 *Let* $P = \{P_1, ..., P_\tau\}$ *be an arbitrary partition of the set of* N *variables, and let* $\Pi = \{P\}$ *be the set of all possible partitions of* N.

In the case of proper simple games considered here, an arbitrary partition consists of a set of winning coalitions (and its complement). Assume, like in [10], that there exists a *partition function* F that assigns a real binary number $F_P(P_i)$ to each coalition $P_i \in P$ for each $P \in \Pi$.

In a simple game, the partition function divides the sets of coalitions in a partition into sets of winning coalitions (W_i) and losing coalitions (L_i). In what follows it will be shown that the vector of *weights* of users coincides with the separating hyperplane between the set of winning coalitions and losing coalitions.

Definition 7 *First define a* $N \times M$ *matrix of winning coalitions* $\bar{\mathbf{G}}$ *such that the rows of* $\bar{\mathbf{G}}$ *denote the various winning coalitions (obtained from the partition function)* $\mathbf{x}=(x_i, ..., x_N)$, *with* $x_i = 1$ *iff* $i \in W_j$, *where* W_j, $j = 1..M$ *is winning coalition indexed* j *(and* $x_i = 0$ *otherwise).*

A simple game $G = (N, W)$ *is said to be weighted if there exists a weight function* $\mathbf{w} : N \rightarrow R$ *and a real number, quota, denoted by* $q \in R$ *such that a coalition is winning precisely when the sum of the weights of the players in* X *exceeds or meets the quota,* $\mathbf{x}'\mathbf{w} \geq q$.

Typically the quota and the weightings are not unique. Here, however, one seeks to find a unique threshold q and $\bar{\alpha}$ as function of q such that capacity for sets of users exceeding this general threshold is always max-min fair (in the temporally defined model the delay cost must be taken into account).

Define a noncooperative game as follows. There are two players, weight selector, choosing a vector of weightings $\mathbf{w} = (w_1, ..., w_N)$, in order to maximize, for each coalition (c.f. fairness), the minimum value that can be quaranteed. Opposed to the weight selector, there is a state selector, nature, choosing a probability distribution $\mathbf{p}$ over the various realizations, to minimize the expected value of any coalition.

Clearly, the game thus defined is a von Neumann minimax game, assuming $\sum p_i = 1$ and $\sum w_i = 1$. Thus, the game has unique value given by

$$q = \min_p \max_{\mathbf{w}} \mathbf{p}'\bar{\mathbf{G}}\mathbf{w} = \bar{\mathbf{p}}'\bar{\mathbf{G}}\bar{\mathbf{w}}. \tag{19}$$

Proposition 8 *In a (self-dual) coherent structure there exist weightings and a quota such that the weightings are max-min fair and the quota is unique.*

The model defines a threshold capacity that is max-min fair:

Proposition 9 *The model yields an expected value (in terms of minimax probabilities) of a completely defined user capacity in the symmetric case given by* $\frac{1}{\alpha^{**}} = \bar{\mathbf{p}}\bar{\mathbf{W}}$, *where* $\bar{\mathbf{W}} = \bar{\mathbf{G}}\mathbf{1}$ *and* $\mathbf{1}$ *is a vector of ones. The optimal threshold capacity is thus given by*

$$\bar{\alpha} = \frac{1}{\bar{\mathbf{p}}'\bar{\mathbf{W}}}, \tag{20}$$

Proof. $\bar{\mathbf{G}}\mathbf{1}$ gives the cardinality vector of the winning coalitions and $\bar{\mathbf{p}}$ is the max-min optimal (equilibrium) probability vector that nature assignes to the cardinality vector. The optimal threshold capacity is the inverse: $\bar{\alpha} = \frac{1}{\bar{\mathbf{p}}'\bar{\mathbf{W}}}$. This threshold extends the definition of max-min fairness to games not generated, i.e. to games that may have various different probability distributions. These games correspond to a sequence (in time/space) of resource allocation problems in a situation where the distribution functions of the random variables are not completely known and can be randomly modified in time.

Remark 1 *The expected value of the completely defined user capacity* $\frac{1}{\bar{\alpha}}$ *is by definition the reliability function of the structure (completely defined boolean description): the reliability function of a completely defined boolean mapping is the function [14]* $\hat{F} : [0,1]^n -> [0,1]$ *(c.f. the definition of a completely defined user capacity):*

$$\hat{F} = Prob\{F = 1\} = E(F) \tag{21}$$

4 Optimal (greedy) algorithm for the centralized model

In what follows we will prove the optimality of the greedy algorithm in solving the centralized resource allocation problem in a multimedia network under incomplete information.

Proposition 10 *The greedy algorithm is optimal in solving resource allocation of the multimedia network.*

Proof. By section *II*, the centralized optimal solution is where the interference from other users is the same for all users: by the proof of *Proposition 2*, the noise from other users is $K = M - X \; \forall i$ (with $K = n - 1$ when $X = \frac{c}{\sum \frac{1}{g_i}} = 1$, for $c = \sum \frac{1}{g_i}$).

Let $S^j = \{1, ..., j\}$ and assume $g_1 \geq g_2 \geq ... \geq g_k > \epsilon \geq g_{k+1} \geq ... \geq g_n, \epsilon \geq 0$. Then, the centralized capacity allocation problem can be written as a linear optimization problem over a polymatroid:

$$max \sum g_i x_i \;\; s.t. \sum_{i \in S} x_i \leq f(S), \tag{22}$$

where f is a submodular function $f : 2^n \to R$ defined by

$$f(S^j) = \sum_{i=1}^{j} x_i^*. \tag{23}$$

By [12] the optimal solution consists of a primal optimum value given by:

$$\sum_{j=1}^{k} g_j(f(S^j) - f(S^{j-1})). \tag{24}$$

By the definition of f the optimum value becomes equal to the sum of received optimal powers: $\sum_{j=1}^{k} g_j x_j^* = k$. Now impose the k-fold noise from others taken as constant K to begin with, by dividing the previous expression by K; this gives the optimal capacity α^* as desired:

$$\frac{kX}{kK} = \frac{X}{K} = \frac{X}{M - X} = \alpha^*. \tag{25}$$

Finally, show that the interference from other users is in fact minimized by maximizing sum of capacities; the dual to problem of maximizing capacities can be written as

$$min\{\sum_{S \subseteq n} f(S) y_S : \sum_{j \in S} y_S \geq g_j,\ y \in R^{2n}\}. \tag{26}$$

The dual objective value is $\sum_{j=1}^{k-1}(g_j - g_{j+1}) f(S^j) + g_k f(S^k) = \sum_{j=1}^{k} g_j x_j^*$, which for given sequence $\{g_j\}$ minimizes the sum of transmit powers.

The optimality of the greedy algorithm follows from the separation between optimal power control and channel allocation. By *Proposition 3* $\alpha^* = \frac{1}{n-1}$ and $\mathbf{AX} \leq \frac{1}{\alpha}$ is totally dual integral, with $x_i^* = \frac{1}{g_i}$ (setting $c = \frac{1}{\sum \frac{1}{g_i}}$). By [12] the system $\sum x_i \leq f(S)$ is TDI and with $x_i^* = \frac{1}{g_i^*}$ (where $\frac{1}{g_i^*} = \lfloor \frac{1}{g_i} \rfloor$) corresponds to the system $\mathbf{AX} \leq \frac{1}{\alpha}$. Thus, given the optimal channel allocation based on the greedy algorithm, the optimal allocation of powers can be done separately at each receiver. Optimal channel allocation rule only requires the existence of a solution to the optimal power allocation which is guaranteed by the von Neumann model, for nonnegative matrices $\mathbf{G}$ and $\mathbf{F}$.

The scheduling rule implied by the greedy algorithm seems similar to that in [4]: a user with higher fading coefficient has priority. An alternative interior point alogrithm to solve the von Neumann problem is presented in [16].

5 Conclusion

This paper has studied resource allocation in a multimedia network. An alternative interpretation of the model is to consider a multireceiver network. Two definitions of max-min fair optimal capacity applicable in both cases were given. First, under partial information inference was based on a set of generated games. Second, a completely defined threshold function was presented. Finally, the optimal algorithm solving the allocation of optimal transmit powers to users, is shown to be the greedy algorithm.

References

[1] S. Anderson, A. de Palma and J.-F. Thisse, *Discrete Choice Theory of Product Differentation*, MIT Press, 1996

[2] T. Basar and P. Bernhard, *H^∞-Optimal control and related minimax design problems*, Birkhauser, 1991

[3] E. Grilich, M. Kovalev and A. Zaporozhets, "A Polynomial Algorithm for Resource Allocation Problems with Polymatroidal Constraints", Optimization, 1996, Vol. 37, pp. 73-86

[4] S. Hanly and D. Tse, "Multi-access fading channels: delay-limited capacities", *IEEE International Symposium of Information Theory*, Camridge 1998, forthcoming in Transactions on Information Theory

[5] T. Heikkinen, "Dynamic Pricing and Power Control of a Multi-Media Network", *Proc. of the Conference in Information Sciences and Systems*, 1998

[6] F. Kelly, "Charging and Rate Control for Elastic Traffic", *European Transactions on Telecommunications*, Vol 9, 1997

[7] G. Koehler, A. Whinston and G. Wright *Optimization over Leontief substitution systems*, North-Holland, 1975

[8] K. Lancaster, *Mathematical Economics*, Macmillan 1968

[9] A. Lazar and N. Semret, A Resource Allocation Game with Application to Wireless Spectrum, Technical Report, Columbia University

[10] W. Lucas, "A Game in Partition Function Form with no Solution", *SIAM J. Appl. Math.*, vol 16, 1968

[11] A. Mas-Colell, M. Whinston, J. Green, *Microeconomic Theory*, Oxford University Press, 1995

[12] G. Nemhauser and L. Wolsey, *Integer and Combinatorial Optimization*, Wiley 1988

[13] A. Prekopa, *Stochastic Programming*, Akademiai Kiado, Budapest 1991

[14] K. Ramamurthy, *Coherent Structures and Simple Games*, Kluwer, 1991

[15] A. Schrijver *Linear and Integer Programming*, Wiley, 1986

[16] Y. Ye, *Interior Point Algorithms, Theory and Practice*, Wiley 1997

[17] J. Zander, "Performance of Optimum Transmitter Power Control in Cellular Radio Systems", *IEEE Transactions on Vehicular Technology*, Vol. 41, No. 1, Feb. 1992

Mitigating the effects of uncertainty in Join-the-Shortest-Queue resource allocation schemes through Prioritized Dispersal

Yitzhak Birk and Noam Bloch
Electrical Engineering Department
Technion — Israel Inst. of Technology
Haifa 32000, Israel
birk@ee, bloch@tx.technion.ac.il

Abstract

Balancing the load among multiple resources is critical for their efficient utilization. It has been shown in various contexts that very good balancing is achieved by assigning each arriving task to the resource with the shortest queue among all queues or even among a randomly chosen small subset of the queues. There are, however, factors that limit the performance of such JSQ schemes in real systems; these include untimely state information and the unknown service time of the queued tasks. In the absence of any queue-state information, a different approach, Prioritized Dispersal, can be effective. There, more resources are requested than are needed, but the extra requests are assigned a low priority; this represents selective exploitation of redundancy based on partial but current information. In this paper, we present a combined scheme, JSQ-PD, whereby the original requests are assigned a high priority and submitted to the (supposedly) shortest queue(s), and redundant low-priority requests are submitted to other queues. Analysis of the case of duplicate requests shows that the new hybrid scheme substantially outperforms JSQ across the reliability range of the state information. Possible applications of this approach are numerous, including parallel computers, storage systems, operating systems and communication networks.

1 Introduction

An attractive way to increase the **potential** performance of a system is to use parallel architectures. By so doing, one can take advantage of the most cost-effective technology; moreover, fault-tolerance is facilitated. However, the potential performance can be fully realized only if the load is equally apportioned among the multiple resources. These may be stateless, such as processors and communication paths in packet-switching networks, so any resource can be used for any task; in other cases, such as storage devices, memory banks and circuit-switched networks, the resources are stateful, in which case the choice among them is limited. The issue of load balancing has been studied extensively for a variety of applications.

Intuitively, the load is balanced by allocating tasks to unloaded resources or by transferring tasks from loaded resources to less loaded ones. In many cases, however, transferring tasks is costly or even impossible. The focus of this paper is on allocation schemes without reallocation.

1.1 "Join the shortest queue" (JSQ) policies

In its purest form, this scheme entails assigning an arriving task to the resource with the least number of tasks in its queue. In practice, the choice may be restricted by the fact that not every resource can carry out every task; examples include the choice of a cell site by a mobile communication unit [1] and mirrored disk arrays. Moreover, even if this restriction does not apply, the cost of discovering the queue length of all resources may be prohibitive, leading to a voluntary restriction on the choice. Mitzenmacher [2] and Vvedenskaya [3] showed that even such limited load balancing suffices for obtaining an exponential improvement in the decay of the tail of the queue-length probability density function when compared with the assignment of a task to a randomly-chosen resource.

Another version of JSQ-based load balancing was presented in [4] and [5] for redundant disk arrays: at recording time, a block of data is partitioned into m subblocks, and r redundant ones are computed from those, such that any m of the $m + r$ suffice for reconstruction [6]. These are then stored in $m + r$ different ("randomly" chosen) disk drives. When a block is requested, any m of the $m + r$ disks containing the subblocks may be accessed. The choice of r for any given value of m represents a trade-off between (storage) overhead and flexibility. [4] further addresses the cost of accessing "redundant" blocks instead of the original ones in the context of a video-on-demand server. It introduces *selective exploitation of redundancy*, which weighs the benefit of load balancing against the extra cost (processing and buffer space) when deciding which subblocks to read.

JSQ has been proposed for numerous applications, including multiprocessor systems [7][8][9], disk arrays [5][4], and communication networks [10] [11]. While it is a good technique, several factors may limit JSQ's effectiveness in real-life situations:

- The queue-length information may be outdated due to physical limitations or the prohibitive overhead associated with constantly updating it (e.g., [12]).
- The service time for each queued task is often unknown. In this case, shortest queue does not necessarily imply least-loaded resource.
- A resource providing service may be stalled, awaiting information from other resources, so even the net service time for a task is not necessarily an accurate measure.

1.2 Prioritized Dispersal [13]

In communication networks that provide multiple paths between a source and a destination, the state of the resources (network links) is never fully known. Maxemchuk [14] proposed *dispersity routing*, whereby a message is partitioned into several (m) submessages, which are then sent to the destination over different paths in a store-and-forward network. Dispersal assists in balancing the load among paths. Also, the different submessages can be transmitted and propagated concurrently, thereby reducing latency.

However, all submessages must be received before the message can be reconstructed, so the delay along the slowest path determines the delay of the message.

To overcome the "weakest link" sensitivity of dispersity routing, Maxemchuk also suggested *redundant dispersity routing*: a number (r) of additional, "redundant" submessages are derived from the original ones and are also transmitted, and any sufficiently large subset of submessages suffices for the reconstruction of the original message. Later, Rabin [15] suggested the *Information Dispersal Algorithm (IDA)*, which implements redundant dispersal, and analyzed it in the context of interconnection networks such as a hypercube.

Redundant dispersal can tolerate some loaded paths, but increases the overall load, reduces capacity and may thus actually increase delay and loss probability. This problem can be mitigated by using the recently-proposed *Prioritized Dispersal* scheme [13], whereby "redundant" submessages receive lower priority than the "original" ones and do not interfere with them; moreover, the use of non-FCFS queuing policies for the redundant submessages leads to the timely arrival of at least a fraction of them even under heavy load.

Casting Prioritized Dispersal in generic terms, it entails partitioning a task into m subtasks along with r redundant subtasks, and assigning the tasks to a randomly-chosen set of $m + r$ resources. While no knowledge pertaining to the load on the resources is available at assignment time, the assignment of low priority to the redundant subtasks acts at each resource to exploit it when available while deferring to "original" subtasks when it is loaded. This helps balance the load among the resources with a bias toward fair treatment of the different tasks. Clearly, this scheme is imprecise, but the partial information that does get used (implicitly) is current, and analysis shows it to offer substantial advantages non-prioritized schemes.

JSQ and Prioritized Dispersal represent different approaches to helping balance the load. The main contribution of this paper is JSQ-PD, a family of schemes that combine the two approaches. Using a simple member of the family, we show that this hybrid approach may offer substantial benefits.

The remainder of the paper is organized as follows. In section 2, we describe the JSQ-PD family of schemes. In section 3, we describe and analyze a specific scheme. Section 4 presents a numerical comparison among schemes, and section 5 offers concluding remarks.

2 The JSQ-PD family of schemes

"Join the Shortest Queue" (JSQ) exploits state information, but must do so at an early stage (task arrival time). Prioritized Dispersal (PD), in contrast, employs an initially random decision; however, the additional low-priority tasks that are assigned to resources permit it to take advantage of those when they are available (without causing any harm). This "decision" whether or not a low-priority task receives service, albeit distributed, imprecise and implicit, is postponed and thus benefits from more current state information at the resources.

Motivated by this observation, we propose to combine the two schemes: the initial decision is based on the available state information, but redundant, low-priority tasks are assigned to additional, presumably busier, servers. The simplest form is task replication, but the same idea can be applied in conjunction with general error-correcting codes. We next describe a specific JSQ-PD scheme, present a queuing model and use it to analyze the scheme's performance in terms of delay distribution.

3 Analysis of a simple JSQ-PD scheme

3.1 The (1,1)JSQ-PD Scheme

For each arriving task, two resources are picked at random (from among the available resources) and considered. One copy of the task is assigned a high priority and is submitted to the resource whose queue is believed to be shorter, and a second copy is assigned a low priority and submitted to the resource whose queue is believed to be longer. In the case of equality or lack of queue-state information, the assignment is random. "$(1, 1)$" stands for 1 "original" task and 1 redundant one. Service is provided by a resource based on priority; the service of an `L-P` task is preempted by an arriving `H-P` task and is eventually resumed. `H-P` tasks are served FCFS, and `L-P` ones are served LCFS.

3.2 Queuing model

We model the set of parallel resources as a system of N parallel queues, one per resource. The two copies of a task are assigned to two randomly chosen queues. N is assumed to be large and the queues are assumed to be i.i.d., so it suffices to analyze a single queue.

The order of service in the queues is FCFS for the high-priority tasks and LCFS for the low-priority ones. A preemptive-resume priority discipline is used to favor `H-P` tasks, and is also applied among low priority tasks (for simplicity of analysis).

The arrival process of `H-P` tasks to any single queue is Poisson with rate λ; task service time is distributed exponentially and independently from task to task, with mean $\frac{1}{\mu}$; the service times of different copies of the same task are also assumed to be independent of one another. The resulting load is $\rho \equiv \frac{\lambda}{\mu}$. The same holds for the `L-P` tasks. We also assume that the `H-P` and `L-P` arrival processes to any given queue are independent.

We assume that the task-assignment algorithm does not know whether the state information that it possesses is reliable. We express this questionable reliability as a probability: for any given arriving task, there is probability P_{lb} that the information is reliable, resulting in a correct JSQ decision; with probability $P_{nb} = 1 - P_{lb}$, the information is unreliable (or unavailable) and the resulting assignment is effectively random.

We now proceed to analyze a single priority queue with FCFS service for `H-P` customers and LCFS service for `L-P` customers. In view of the symmetry and independence among queues, this suffices. It should, however, be noted that the queue states as seen by arriving tasks can become coupled through the assignment policy. With pure JSQ, for example, it is known that the queue joined by the `L-P` task is at least as long as that joined by the corresponding `H-P` task. This coupling is captured by the analysis. We begin the analysis by computing occupancy probabilities, and then use those to compute the distribution of delay.

3.3 Occupancy probabilities for the $(1, 1)$-JSQ-PD scheme

Vvedenskaya, Dobrushin and Karpelevich [3] and Mitzenmacher [2] considered a system in which tasks are assigned to the resource with the shortest queue among d resources chosen randomly from among $N >> d$ resources. They do not assume that queues are independent (but the results obtained are asymptotically the same as those obtained for independent queues). They analyzed such a system with Poisson arrivals of tasks and exponentially distributed service times, and showed that the decay of the tail of the occupancy probability function is doubly exponential. This is an exponential improvement over the non-balanced case, wherein each task is assigned to a randomly chosen server. In [16], the analysis was extended to the case wherein reliable queue-length information exists and is used with probability P_{lb}; with probability $P_{nb} = 1 - P_{lb}$, such information is unavailable and the assignment is made randomly. (We use "nb" and "lb" to denote no balancing (random assignment) and load balancing (JSQ), respectively.)

Our analysis of (1, 1)JSQ-PD uses the fact that L-P tasks do not interfere with the H-P ones. Consequently, the occupancy probabilities for H-P tasks are equal to those derived by [16] for a single priority level.

Let L_i be the probability for at least i high-priority tasks in a randomly chosen queue. By [2],

$$L_i = \rho L_{i-1}\left((1 - P_{lb}) + (P_{lb})L_{i-1}\right); \quad L_0 = 1, \qquad 0 \le p_{lb} \le 1. \tag{1}$$

For pure JSQ (reliable information always available), this becomes

$$L_i = \rho^{2^i - 1}, \qquad p_{nb} = 0. \tag{2}$$

P_i, the probability for i high-priority tasks in a randomly chosen queue, is simply

$$P_i = L_i - L_{i+1}. \tag{3}$$

Let $P^H(k|lb)$ and $P^H(k|nb)$ denote the probability that an arriving H-P task joins a queue with k H-P customers given that it chose the shorter of the two queues or a queue at random, respectively. Also, let $P^L(i|lb, k)$ and $P^L(i|nb, k)$ denote the probability that an arriving L-P task joins a queue with i H-P tasks given that its H-P partner, having chosen the shorter of two queues or one of them at random, respectively, joined a queue with k H-P customers.

From a priori symmetry considerations, it follows that when the queues are chosen at random,

$$P^H(k|nb) = P_k, \quad k \ge 0 \tag{4}$$

and

$$P^L(i|nb, k) = P_i, \quad i \ge 0. \tag{5}$$

When the H-P task joins the shortest queue,

$$P^H(k|lb) = 2 \cdot P_k \left(1 - \sum_{j=0}^{k} P_j\right) + P_k^2, \quad k \ge 0 \tag{6}$$

and

$$P^L(i|lb, k) = \frac{P_i}{\sum_{j=k}^{\infty} P_j}, \quad i \ge k. \tag{7}$$

3.4 Delay Distribution

The delay incurred by a task is the minimum of the delays incurred by its copies. In this section we derive $D(t)$, the probability that this delay is smaller than t. (Delay is expressed in units of mean task service time.) We first derive, for the H-P and L-P copy, the probability that its delay is smaller than t, given that the H-P task sees k H-P tasks upon arrival. We do this separately for the cases of reliable (lb) and unreliable (nb) information. Then, we remove the conditions by summing over k.

The delay of the H-P task consists of the service times of the k tasks that it finds in the queue as well as its own service time. With an LCFS–preemptive-resume queuing discipline for the L-P tasks, the delay of a tagged L-P task depends only on the number of H-P tasks that it finds upon arrival and on the number of H-P and L-P tasks that arrive (subsequently) before its service ends. The number of L-P tasks found in the queue at the tagged L-P task's arrival time does not affect its delay because they are not served until the tagged task's service ends.

Let $D^H(t|k,nb)$ and $D^H(t|k,lb)$ denote the probability that the delay of an H-P task is smaller than t given that it joins a queue with k H-P tasks based on JSQ or random assignment, respectively. Also, let $D^L(t|k,nb)$ and $D^L(t|k,lb)$ denote the probability that the delay of an L-P task is smaller than t given that its H-P partner joins a queue with k H-P tasks based on JSQ or random assignment, respectively. Then,

$$D(t) = P_{nb} \sum_{k=0}^{\infty} P^H(k|nb) \cdot D(t|k,nb) + P_{lb} \sum_{k=0}^{\infty} P^H(k|lb) \cdot D(t|k,lb), \text{ where} \tag{8}$$

$$D(t|k,nb) = 1 - (1 - D^H(t|k,nb)) \cdot (1 - D^L(t|k,nb)), \text{ and} \tag{9}$$

$$D(t|k,lb) = 1 - (1 - D^H(t|k,lb)) \cdot (1 - D^L(t|k,lb)). \tag{10}$$

The delay of an H-P task that joins a queue with k H-P tasks consists of $k+1$ independent, exponentially distributed service times, so

$$D^H(t|k,lb) = D^H(t|k,nb) = \int_0^t \frac{\mu(\mu x)^k \exp^{-\mu x}}{k!} dx. \tag{11}$$

If the H-P task joins a queue with k H-P tasks based on JSQ, its partner L-P task joins a queue with at least k H-P tasks. In the case of a random selection, the arriving L-P task sees the a priori queue length distribution. Consequently,

$$D^L(t|k,nb) = \sum_{i=0}^{\infty} D_i^L(t) P_i \text{ and} \tag{12}$$

$$D^L(t|k,lb) = \sum_{i=k}^{\infty} D_i^L(t) \cdot P^L(i|lb,k). \tag{13}$$

Substituting for $P^L(i|lb,k)$ from (7) yields

$$D^L(t|k,lb) = \sum_{i=k}^{\infty} D_i^L(t) \frac{P_i}{1 - \sum_{j=0}^{k-1} P_j}, \tag{14}$$

where $D_i^L(t)$ is the probability that an arriving L-P task that finds i H-P tasks in the queue incurs a delay less than or equal to t.

Derivation of $D_i^L(t)$. To derive the sojourn time (delay) distribution of a low-priority task that finds i high-priority tasks in the queue when it arrives, we first define a stochastic process $U(t)$, the remaining work (at time t) that should be done before the termination of the service of the tagged low priority task that arrived at time t_0. The sojourn time of the tagged task is the time needed for U to reach zero. $U(t_0^-)$ consists of the residual service time of the H-P task in service plus the service time of the high priority tasks waiting in the queue. $U(t_0)$ includes, in addition, the tagged task's service time. If high- or low-priority tasks arrive before U reaches zero, their service times are added to U. One may notice that, since the system is work conserving, $U(t)$ is independent of the order in which service is granted to the tasks that are served prior to the completion of the tagged L-P task's service.

For convenience, we assume the following order of service: first serve the H-P task in service, the high priority tasks that were waiting in the queue at t_0 and the tagged L-P task. This will be considered the "first service", with Laplace transformed time distribution $B_f^*(s)$. Then serve the (high and low priority) tasks that arrive during the first service and afterward until U reaches zero.

This procedure is the same as the one used in deriving the distribution of busy period duration in a queue in which there is exceptional first service ($U(t_0) = B_f^*(s)$ in our case) and exponentially distributed subsequent service times. For such systems [17], the Laplace transform of the busy period is

$$D_{i_l}^*(s) = B_f^*(s + 2\lambda - 2\lambda Y_0^*(s)), \tag{15}$$

where $Y_0^*(s)$ is the Laplace transform of the busy period in an ordinary M/G/1 queue (with non-exceptional first service) and arrival rate 2λ. It can be calculated as

$$Y_0^*(s) = B^*(s + 2\lambda - 2\lambda Y_0^*(s)) \tag{16}$$

and

$$B_{i_f}^*(s) = \left(\frac{\mu}{s+\mu}\right)^{i+1}. \tag{17}$$

Solving for $Y_{0(s)}^*$ yields

$$Y_{0(s)}^* = \frac{1}{4\lambda}\left(\mu + 2\lambda + s - \left((\mu + 2\lambda + s)^2 - 8\lambda\mu\right)^{0.5}\right) \tag{18}$$

Remark. The final steps of the above analysis hinge on the observation that the combined (H-P and L-P tasks) arrival process to a queue is Poisson(2λ), since the two resources for any given task are chosen at random and each is assigned a copy of the task. Thus, with an LCFS queuing discipline for L-P tasks, the component of the delay incurred by an L-P task due to later-arriving tasks can be calculated with relative ease.

Unfortunately, it is impossible to obtain an explicit expression for $D_i^L(t)$ and therefore also for $D(t)$. Numerical results for $D_i^L(t)$ are obtained through a numerical inverse Laplace transform. Moreover, in some cases, summation over an infinite number of elements is required; instead, we sum up only a finite number of elements, the number of which depends on the desired accuracy.

4 Numerical Results and Comparison

In this section, we present a comparison among JSQ, PD and JSQ-PD. For JSQ, two randomly chosen resources are considered and a task is assigned to the one with the supposedly shorter queue. With PD, a single high-priority task and a single low-priority one are assigned randomly to two randomly-chosen resources. With JSQ-PD, two randomly-selected resources are considered: the one with the supposedly shorter queue receives a high-priority task, and the other receives its low-priority copy. Service time of tasks is distributed exponentially, and the arrival process of tasks to the system is Poisson.

Fig. 1 depicts the probability of delay exceeding T times the mean task service time. On the left, $T = 3$ and the probability is plotted as a function of load ρ with fixed information-reliability $P_{lb} = 0.5$; on the right, it is plotted as a function of the threshold

T with fixed $P_{lb} = 0.5$ and $\rho = 0.7$. It can be seen that JSQ-PD outperforms the other schemes across the entire range; this is to be expected since there is no penalty for the redundant (low-priority) tasks. The advantage of JSQ-PD over JSQ is more pronounced with a lower load, since with a very high load the low-priority task is very unlikely to find an unloaded resource and be served earlier than its high-priority partner. (With $\rho = 0.9$, for example, only a ninth of the L-P tasks are ever served.) The improvement is almost independent of T.

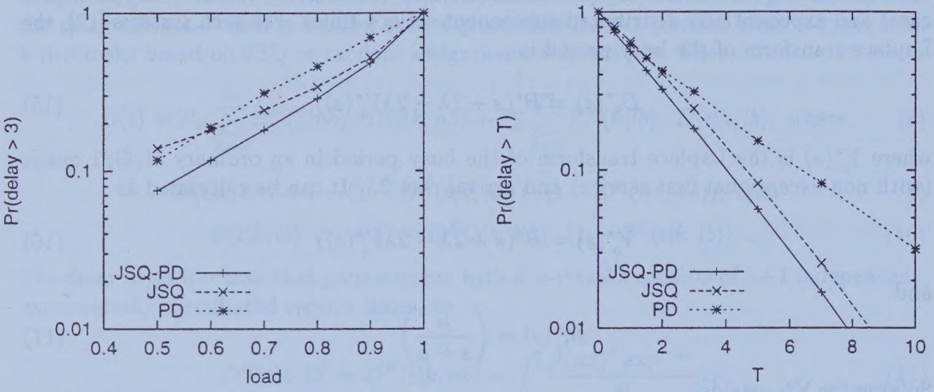


Figure 1: The probability for the delay to exceed T mean task service times as a function of the load (left) and T (right); $P_{lb} = 0.5$.

Fig. 2 depicts the probability of delay exceeding three and ten times the mean task service time as a function of the reliability of the queue-state information, expressed by P_{lb}. Results are presented for two levels of load, $\rho = 0.5$ and $\rho = 0.9$.

It can be seen that JSQ-PD outperforms the other schemes across the entire range of P_{lb}; The performance of Prioritized Dispersal is independent of the reliability of state information since such information is not used by this scheme. Its performance, in fact, is equal to that of JSQ-PD with $P_{lb} = 0$. As expected, the relative performance of JSQ and PD depends on P_{lb}. Surprisingly, Fig. 2 suggests 1) an apparent exponential dependence of the probability for the delay to exceed a threshold on P_{lb}, and 2) an apparent independence of the relative advantage of JSQ-PD over JSQ from the value of P_{lb} for an exponential service time. Both observations are quite accurate (albeit not precise) for various values of load and delay thresholds. The first one may indicate that there is an exponential approximation for the dependence of delay on P_{lb}.

A similar comparison for a fixed service time, conducted via simulation, shows a smaller advantage to JSQ-PD over JSQ. This can be explained by the fact that one of the uncertainty factors, service times, is eliminated (except for the residual service time of the H-P task being served when the tagged task joins the queue). For the same reason, the two are very close as P_{lb} approaches one.

The model used for the analysis is obviously an approximation. However, having verified the closeness of the independence approximation by simulation and in view of the substantial performance differences among the schemes, we conjecture that the results provide a good indication of the merits of JSQ-PD.

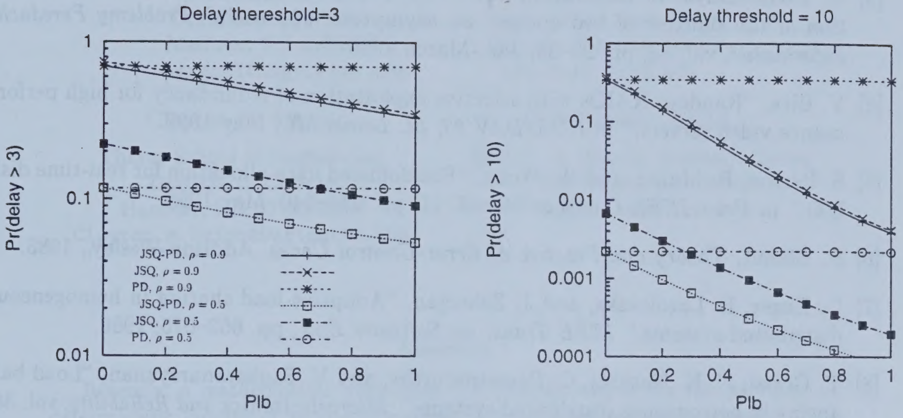


Figure 2: The probability for the delay to exceed 3 (left) and 10 (right) mean task service times as a function of P_{lb}.

5 Conclusions

In this paper, we combined the well-known "Join the Shortest Queue" family of schemes with the recently-proposed Prioritized Dispersal schemes. The former improve load balancing among multiple resources based on knowledge of their state and do so in a direct way. The latter, in contrast, generate redundant tasks and submit them with low priority, thus soliciting more work than is required but never interfering with the execution of the "original", high-priority tasks. Thus, they can take advantage of available resources to overcome overloading in others. This indirect, imprecise approach benefits from the fact that the decision whether to exploit resources is partly deferred and distributed to a location and time at which state is known. The two approaches are complementary, and the analysis of a hybrid scheme whereby two copies of each task are submitted, using any available knowledge to try and submit the high-priority copy to the less loaded resource, indeed shows significant improvements over the constituent schemes.

The hybrid schemes, dubbed JSQ-PD, are applicable to most situations in which the constituent schemes are used, and consideration of their use is highly recommended along with further research aimed at improving them as well as analyzing the performance of more complicated family members.

References

[1] M. Alanyali and B. Hajek, "On simple algorithms for dynamic load balancing," in *Proc. IEEE INFOCOM'95*, pp. 230–238, 1995.

[2] M. Mitzenmacher, "On the analysis of randomized load balancing schemes," in *SPAA '97*, pp. 292–301, June 1997.

[3] N. Vvedenskaya, R. Dobrushin, and F. Karpelevich, "Queueing system with selection of the shortest of two queues: an asymptotic approach," *Problemy Peredachi Informatsii*, vol. 32, pp. 20–34, Jan.-March 1996.

[4] Y. Birk, "Random RAIDs with selective exploitation of redundancy for high performance video servers," in *NOSSDAV'97, St. Louis, MO*, May 1997.

[5] S. Berson, R. Muntz, and W. Wong, "Randomized data allocation for real-time disk I/O," in *Proc. IEEE Compcon'96*, vol. 11, pp. 631–640, May 1996.

[6] R. Blahut, *Theory and Practice of Error-Control Codes.* Addison-Wesley,, 1983.

[7] D. Eager, E. Lazokwska, and J. Zahorjan, "Adaptive load sharing in homogeneous distributed systems," *IEEE Trans. on Software Eng.*, pp. 662–675, 1986.

[8] T. Gopal, N. K. Nataraj, C. Pamamaturthy, and V. Sankaranarayanan, "Load balancing in hetrogenous distributed systems," *Microelectronics and Reliability*, vol. 36, no. 9, pp. 1279–1286, 1996.

[9] M. Willebeck-LeMair and A. Reeves, "Strategies for dynamic load balancing in highly parallel computers," *IEEE Trans. Parallel and Distributed Systems*, vol. 4, no. 9, 1993.

[10] R. Gibbens, F. Kelly, and S. Turner, "Dynamic routing in multiparented networks," *IEEE Trans. Networking*, vol. 1, pp. 261–270, April 1993.

[11] T. Ko and M. Chang, "Least measured cost routing in VP-based ATM networks," in *GLOBECOM 97*, vol. 3, pp. 1820–1824, 1997.

[12] I. Philip, "Dynamic load balancing in distributed systems," in *SOUTHEASTCON'90*, vol. 1, pp. 304–7, 1990.

[13] Y. Birk and N. Bloch, "Improving performance of multi-path ATM networks via prioritized dispersal," Submmited for publication.

[14] N. Maxemchuk, "Dispersity routing," in *Proc. Int. Commun. Conf.*, pp. 41.10–41.13, 1975.

[15] M. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance," *J. ACM*, vol. 36, pp. 335–3488, April 1989.

[16] M. Mitzenmacher, *The power of two choices.* PhD thesis, University of California, Berkeley, 1995.

[17] L. Kleinrock, *Queueing Systems*, vol. 1. Wiley, 1976.

A Framework for Adaptive Service Guarantees*

Rajeev Agrawal
ECE Department
University of Wisconsin
Madison, WI 53706-1691
agrawal@engr.wisc.edu

R. L. Cruz
ECE Department
Univ. of Cal., San Diego
La Jolla, CA 92093-0407
cruz@ece.ucsd.edu

Clayton M. Okino
Thayer School of Engineering
Dartmouth College
Hanover, NH 03755-8000
Clayton.M.Okino@Dartmouth.EDU

Rajendran Rajan
IBM T. J. Watson Research Center
Hawthorne, NY 10532
raju@watson.ibm.com

Abstract

Recently, a mathematical framework has been developed for the provision of deterministic quality of service guarantees in integrated services networks. This framework, or so-called "network calculus," involves the concepts of traffic envelopes, service curves, and convolution in the min-plus alegebra. Traffic envelopes constrain arrival processes, while service curves constrain the input-output behavior of network elements. Upper bounds on network delay are implied by the distance between a traffic envelope and service curve.

In this paper, we develop a somewhat parallel framework for the provision of deterministic quality of service guarantees to *adaptive* applications. Adaptive applications generate a traffic load that is dependent on network utilization, and thus characterizing the traffic generated from an adaptive application with an envelope is problematic. We introduce an adaptive service definition, through which upper bounds on network delay can be derived without using a traffic envelope. Instead, upper bounds on network delay are obtained in terms of the backlog and an *absolute* service curve. Since the backlog can be controlled through feedback, e.g. through window flow control, this yields a mechanism to obtain upper bounds on network delay. Lower bounds on network throughput are also implied through an absolute service curve.

1 Introduction

Adaptive network applications utilize feedback from the network to adjust the rate at which traffic is generated, so that network resources can be used efficiently. When an adaptive traffic source senses that the network is underutilized, it may choose to inject traffic into the network at a higher rate. The traffic may be generated from a fixed size data object, as in a file transfer, where the short term traffic rate is modulated

---

[*] Research supported in part by the National Science Foundation under grants NCR-93-05018, NCR 94-15684, NCR 97-25568, and by the Center for Wireless Communications at UCSD.

in accordance with a window flow control protocol. Alternatively, the total volume of traffic generated may in fact depend on the network feedback, as in adaptive multimedia compression algorithms that encode data with a lower distortion when the network is underutilized. With such adaptive applications, it is difficult to predict in advance the range of traffic behavior. At the same time, such adaptive applications may require quality of service guarantees from the network in order to function. For example, upper bounds on network latency may be required for interactive applications, and a network throughput guarantee may be required as well.

Recently, a mathematical framework has been developed for obtaining deterministic bounds on quality of service in integrated services networks [1][2][3][4][5][7]. Within this framework, upper bounds on delay are obtained with respect to a traffic envelope, which places an upper bound on the amount of data generated over intervals of time. For traffic that can be well characterized in terms of an envelope, bounds on delay, buffering requirements, and throughput can be obtained.

Adaptive applications, however, may generate traffic that is not well characterized by an envelope. Lower bounds on throughput, independent of a traffic envelope, are implied through a service curve [1]. Thus, within this mathematical framework, it is still possible to obtain throughput guarantees for adaptive applications. However, without a traffic envelope, it is difficult to obtain meaningful bounds on delay for an adaptive application that changes its traffic rate in response to network conditions.

In this paper, we develop a new mathematical framework which is useful for modelling services which guarantee upper bounds on delay *and* lower bounds on throughput, without requiring arrival traffic to be constrained by an envelope. We introduce an adaptive service definition which is the foundation of this framework. We shall see that the adaptive service definition is robust in the sense that we may obtain adaptive service guarantees for end-to-end systems that consist of network elements that themselves provide adaptive service guarantees. In particular, we shall obtain adaptive service guarantees for window flow control protocols. We shall also discuss a network element called an "elastic regulator" that can be used to allocate adaptive service guarantees, while allowing a traffic stream to utilize excess bandwidth.

In the next section, we shall review pertinent previous results and concepts, which provide the building blocks and the motivation for the mathematical framework developed in this paper.

2 Previous Results and Concepts

2.1 Notation and Preliminaries

We define a *process* to be a function of time, e.g. $A(t)$. Formally, A is a mapping from the real numbers into the extended non-negative real numbers, i.e. $A : \mathbb{R} \rightarrow \mathbb{R}_+ \cup \{+\infty\}$. A process could count the amount of data arriving or departing to/from some network element, and in this case we may call the process an arrival process or a departure process, respectively. All processes are assumed to be non-decreasing and right continuous. We shall often consider what we call *causal* processes, which are simply processes which are identically zero for all negative times. For example, if A is a causal arrival process to a network element, then $A(t)$ is equal to the amount of data (in bits) arriving to the network element in the interval $(-\infty, t]$, and $A(t) = 0$ for all $t < 0$.

Given two processes A and B, the convolution of A and B is defined to be the function

$A * B : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ such that

$$A * B(t) = \inf_{\tau \in \mathbb{R}} \{A(\tau) + B(t - \tau)\} .$$

It is easy to verify that $A * B$ is a process, i.e. it is non-decreasing, and right continuous. Furthermore, if A and B are causal, then $A * B$ is causal.

The convolution operator is **commutative** and **associative** , i.e. $A * B = B * A$ and $(A * B) * C = A * (B * C)$. Furthermore, if $B \wedge C$ denotes the pointwise minimum of B and C, then $A * (B \wedge C) = (A * B) \wedge (B * C)$. In other words, convolution is **distributive** with respect to the minimum operator. The identity element δ of this operator that satisifes $A * \delta = \delta * A = A$ may be verified to be

$$\delta(t) = \begin{cases} 0 & , t < 0 \\ \infty & , t \geq 0 . \end{cases}$$

Defining $\delta_d(t) := \delta(t - d)$, note that δ_d is a "shift element," i.e. for any process A we have $A * \delta_d(t) = A(t - d)$.

If G is a causal process, then[1] $G \leq \delta$. Thus, if F is any process and G is any causal process, then $F * G \leq F * \delta = F$.

The "deconvolution" of A and B, or A deconvolved with B, is defined by

$$A \oslash B(t) = \sup_{\tau \in \mathbb{R}} \{A(t + \tau) - B(\tau)\}$$

for all t. It is easy to verify that $A \oslash B$ is a process, i.e. it is non-decreasing, and right continuous. Furthermore, if A and B are causal, then $A \oslash B$ is causal. It can be verified that $A \oslash B$ is the smallest process H such that $H * B \geq A$. In other words, if $H * B \geq A$ then $H \geq A \oslash B$, and furthermore $(A \oslash B) * B \geq B$.

We use the notation x^+ to denote $\max\{x, 0\}$.

2.2 Service Guarantees

Suppose that the arrival of traffic to a network element is described by the cumulative arrival process R_{in}. The network element is a *service curve element* with *minimum service curve* S^{min} if the departure process R_{out} from the element satisfies $R_{out} \geq R_{in} * S^{min}$.

A process E is said to be an *envelope* for the process R if for all $\tau \leq t$ we have $R(t) - R(\tau) \leq E(t - \tau)$, or equivalently $R \leq R * E$.

A process E is said to be *sub-additive* if for all $t, \tau \in \mathbb{R}$ we have $E(\tau) + E(t - \tau) \geq E(t)$. Thus, if E is a sub-additive process, then $E * E \geq E$. It is usually assumed that envelopes are sub-additive and causal.

Given any process F, then we use the notation $F^{(\infty)}$ to denote the *sub-additive closure* [3] of F, i.e. $F^{(\infty)} = \wedge_{n=0}^{\infty} F^{(n)}$ and $F^{(n)}$ denotes the n-fold convolution with itself ($F^{(0)} = \delta$).

The backlog B of a network element describes the amount of data stored in the network element as a function of time. If the arrival and departure process of the network element are R_{in} and R_{out}, then the backlog $B(t)$ is given by

$$B(t) = R_{in}(t) - R_{out}(t) ,$$

---

[1] In this paper, all inequalities involving functions are defined in a pointwise sense.

assuming that no data is stored in the network element at time $t = 0$. If data from the network element departs in the same order that it arrives, then data arriving at time t will depart by time $t + \Delta$ only if $R_{out}(t + \Delta) \geq R_{in}(t)$. This motivates the following definition of $d(t)$, the virtual delay at time t:

$$d(t) = \inf\{\Delta : \Delta \geq 0 \text{ and } R_{out}(t + \Delta) \geq R_{in}(t)\} .$$

One of the fundamental previous results concerning envelopes and service curves is that delay can be upper bounded by maximum horizontal distance between an envelope and a service curve. This is formalized in the following proposition.

Proposition 1 *[5][1][7][2] Consider a service curve element with minimum service curve S^{min}. If the arrival process has envelope E, then the virtual delay $d(t)$ satisfies $d(t) \leq d^{max}$ for all t, where*

$$d^{max} = \inf\{d : d \geq 0 \text{ and } E * \delta_d \leq S^{min}\} .$$

It is easy to see that if the arrival process to a service curve element is unconstrained, then the virtual delay can be unbounded, since the backlog may be unbounded. Adaptive applications may use feedback to adjust the arrival process to a network so that the backlog remains bounded. However, even if the backlog of a service curve element remains bounded, the virtual delay may still be unbounded. This is because the condition $R_{out} \geq R_{in} * S^{min}$ is not strong enough to guarantee that R_{out} will increase over an arbitrary interval where the backlog is positive. For example, suppose $R_{out}(s)$ is significantly larger than $R_{in} * S^{min}(s)$, say $R_{out}(s) \geq R_{in} * S^{min}(s + \Delta)$, where Δ is large. In this case, the constraint $R_{out}(u) \geq R_{in} * S^{min}(u)$ can be met for $u \in (s, s + \Delta)$ even when no data departs the network element in the interval $[s, s + \Delta]$, i.e. $R_{out}(s + \Delta) = R_{out}(s)$. In other words, roughly speaking, if a traffic stream receives significantly more service than it is guaranteed up to a time s, then after time s it is possible for a service curve element to serve very little traffic, and still meet the guarantee specified by the service curve. In order to guarantee an adaptive session a bounded delay, it should not be "punished" later for utilizing excess bandwidth at an earlier time. This motivates the adaptive service definition introduced in the next section.

3 Adaptive Service Guarantees

Definition 2 (Adaptive Service Guarantee) *Given a network element with arrival process R_{in} and departure process R_{out}. Let $\tilde{S}$ and S be causal processes. We say the network element adaptively guarantees $(S, \tilde{S})$ over $[s^*, t^*]$, if for all $s, t \in [s^*, t^*]$ and $s \leq t$ we have*

$$R_{out}(t) \geq \{R_{out}(s) + \tilde{S}(t - s)\} \wedge \inf_{u: t \geq u \geq s} \{R_{in}(u) + S(t - u)\} . \tag{1}$$

If (1) holds for all $s, t \in [s^*, t^*]$ and $s \leq t$, then we write, as a shorthand notation,

$$R_{in} \rightarrow (S, \tilde{S})_{[s^*, t^*]} \rightarrow R_{out} .$$

If (1) holds for all $s, t \in \mathbb{R}$ and $s \leq t$, then we simply say that the network element adaptively guarantees $(S, \tilde{S})$, and we write as a shorthand notation

$$R_{in} \rightarrow (S, \tilde{S}) \rightarrow R_{out} .$$

In the context of these definitions, we say that S is a partial service curve, and that $\tilde{S}$ is an absolute service curve.

In the special case where $S = \tilde{S}$ and $R_{in} \rightarrow (S, \tilde{S})_{[s^*,t^*]} \rightarrow R_{out}$, we simply say that the network element adaptively guarantees $\tilde{S}$ over $[s^*, t^*]$, and we write $R_{in} \rightarrow (\tilde{S})_{[s^*,t^*]} \rightarrow R_{out}$. Similarly, if $R_{in} \rightarrow (\tilde{S}, \tilde{S}) \rightarrow R_{out}$ then we simply say that the network element adaptively guarantees $\tilde{S}$, and we write $R_{in} \rightarrow (\tilde{S}) \rightarrow R_{out}$.

To get a rough sense of the meaning of these definitions, suppose that $R_{in} \rightarrow (S, \tilde{S}) \rightarrow R_{out}$, and fix $s \leq t$. Note that the lower bound on $R_{out}(t)$ in (1) potentially increases with $R_{out}(s)$, and it is in this sense that the guarantee is adaptive. Said another way, the amount traffic that departs over (s, t), namely $R_{out}(t) - R_{out}(s)$ is at least $\tilde{S}(t - s)$, assuming that R_{in} is sufficiently large over (s, t), and this is independent of how much traffic departed prior to time s. This also motivates why we have called $\tilde{S}$ an absolute service curve. The partial service curve S influences the lower bound on $R_{out}(t) - R_{out}(s)$ in the case where R_{in} is "small" over the interval (s, t). Other motivations for the terminology will become more clear once we examine systems composed of network elements that provide adaptive service guarantees.

Next, we give two simple examples of network elements which provide adaptive guarantees.

Example 3 (Constant Delay Element) *Consider a network element with constant delay $d \geq 0$, where $R_{out}(t) = R_{in}(t - d)$ for all t. There holds $R_{in} \rightarrow (\delta_d) \rightarrow R_{out}$.*

Example 4 (GPS Server) *Consider a session passing through a Generalized Processor Sharing (GPS) [9] server, with arrival process R_{in} and departure process R_{out}. One of the properties of a GPS server is that each session passing through it has a guaranteed minimum bandwidth, say ϕ. In particular, if $B(u) = R_{in}(u) - R_{out}(u) > 0$ for $u \in (s, t)$, then $R_{out}(t) - R_{out}(s) \geq \phi(t - s)$. Define the process μ_ϕ according to $\mu_\phi(t) = \phi t$ if $t \geq 0$ and $\mu_\phi(t) = 0$ for $t < 0$. It is not difficult to show that the GPS server adaptively guarantees μ_ϕ to the session, i.e. $R_{in} \rightarrow (\mu_\phi) \rightarrow R_{out}$.*

The guarantee provided by GPS motivates the following definition.

Definition 5 (Strong Service Guarantee) *Consider a network element with arrival process R_{in}, departure process R_{out}, and backlog $B = R_{in} - R_{out}$. Let S^* be a process. Suppose the network element has the property that for all intervals (s, t) where* $\inf\{B(u) : u \in [s, t]\} > 0$ *there holds $R_{out}(t) - R_{out}(s) \geq S^*(t - s)$. In this case, we say that the network element strongly guarantees S^*, and we call S^* a strong service curve.*

For example, one of the defining properties of a GPS server is that it strongly guarantees μ_ϕ, in the context of Example 4. In fact, a strong service curve is consistent with the notion of the "universal service curve" originally introduced in [9]. An adaptive service guarantee implies a strong service guarantee, as we now discuss.

Proposition 6 (Strong Service from Adaptive Service Guarantee) *Consider a network element with arrival process R_{in} and departure process R_{out}. If $R_{in} \rightarrow (S, \tilde{S}) \rightarrow R_{out}$, then the network element strongly guarantees S^*, where $S^* = \tilde{S} * S^{(\infty)}$.*

For example, since $\mu_\phi = \mu_\phi * \mu_\phi$, it follows that if a network element adaptively guarantees μ_ϕ, then it stongly guarantees μ_ϕ.

As another example, consider a network element with constant delay $d > 0$. Combining the results of Example 3 and Proposition 6, it follows that a constant delay element strongly guarantees $S^* \equiv 0$. It is clear that, while trivial, this is the best result possible since we may inject an arbitrarily small amount of traffic into the delay element for an arbitrarily long interval of time, while keeping the backlog strictly positive throughout the interval.

The following result illustrates a distinction between an strong service guarantee and an adaptive service guarantee.

Proposition 7 *If $R_{in} \rightarrow (S, \tilde{S}) \rightarrow R_{out}$, and the network element has a backlog at time $s \leq t$ such that $B(s) \geq \tilde{S}(t-s)$, then $R_{out}(t) - R_{out}(s) \geq \tilde{S}(t-s)$.*

For all x, define the "pseudo inverse" of $\tilde{S}$, $\tilde{S}^{-1}$ according to

$$\tilde{S}^{-1}(x) = \inf\{y : y \geq 0, \tilde{S}(y) \geq x\} .$$

Using the right continuity of $\tilde{S}$, it follows that $\tilde{S}(\tilde{S}^{-1}(x)) \geq x$ for all x.

We now present an upper bound on virtual delay in terms of the backlog, which is almost equivalent to Proposition 7.

Proposition 8 (Delay Bound from Backlog) *If $R_{in} \rightarrow (S, \tilde{S}) \rightarrow R_{out}$, then the virtual delay at time t, $d(t)$, is upper bounded according to $d(t) \leq \tilde{S}^{-1}(B(t))$.*

The significance of Proposition 8 is that the delay bound does not require the arrival process R_{in} to be constrained by an envelope, but only that the backlog B is bounded. One possible way to ensure that the backlog is bounded is to employ feedback, as we shall discuss later, within the context of window flow control.

3.1 Composition Results

We first consider a tandem series of n network elements, with no feedback.

Proposition 9 (Composition of Adaptive Service Guarantees) *Suppose for $i = 1, \cdots, n$, we have*

$$R_{i-1} \rightarrow (S_i, \tilde{S}_i)_{[s^-, t^-]} \rightarrow R_i \ .$$

Then

$$R_0 \rightarrow (S_1 * S_2 * \cdots * S_n, \tilde{G})_{[s^-, t^-]} \rightarrow R_n \quad ,$$

where $\tilde{G} = (\tilde{S}_1 * S_2 * \cdots * S_n) \wedge (\tilde{S}_2 * S_3 * \cdots * S_n) \wedge \cdots \wedge (\tilde{S}_{n-1} * S_n) \wedge \tilde{S}_n$.

Next, we consider network elements with feedback.

Consider the closed loop window based flow control model depicted in Figure 1, which consists of network element N_1, network element N_2, and another network element called a *throttle*. The arrival process to the throttle is R_{in}, and the departure process of the throttle, $\hat{R}$, is as large as possible, subject to a "throttle control process" $W + F$. In particular, we have

$$\hat{R}(t) \quad = \quad R_{in}(t) \wedge \{F(t) + W(t)\} \ ,$$

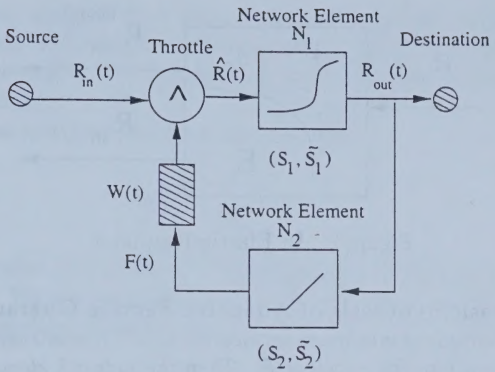


Figure 1: A single hop with window flow control

for all t. Thus, the throttle may store traffic. The throttle departure process feeds network element N_1, whose departure process is denoted as R_{out}. Finally, the arrival process to network element N_2 is R_{out}, and the departure process of N_2 is given by the process F.

We call W the *window size process*, even though it may be a decreasing function. It is assumed, however, that the throttle control process $F + W$ is a process, i.e. it is a non-negative, non-decreasing, right-continuous function.

Note that the total amount of traffic stored in N_1 and N_2 at time t is given by $\hat{R}(t) - F(t) \leq (F(t) + W(t)) - F(t) = W(t)$, and thus $W(t)$ controls the backlog in N_1 and N_2.

Proposition 10 (Adaptive Service Guarantee for Window Flow Control) *If* $\hat{R} \to (S_1, \tilde{S}_1)_{[s^*,t^*]} \to R_{out}$, $R_{out} \to (S_2, \tilde{S}_2)_{[s^*,t^*]} \to F$, $\hat{R} = R_{in} \wedge (F + W)$, *and* $W(u) = w > 0$ *for all* $u \in [s^*, t^*]$, *then* $R_{in} \to (Y, \tilde{Y})_{[s^*,t^*]} \to R_{out}$, *where* $Y = \wedge_{n=0}^{\infty} [S_1 * G^{(n)} + nw]$, $\tilde{Y} = \wedge_{n=0}^{\infty} [\tilde{G} * G^{(n)} + nw]$, $G = S_1 * S_2$, *and* $\tilde{G} = \tilde{S}_1 \wedge (\tilde{S}_2 * S_1)$.

Corollary 11 *Suppose* $\hat{R} \to (S_1, \tilde{S}_1)_{[s^*,t^*]} \to R_{out}$, $R_{out} \to (S_2, \tilde{S}_2)_{[s^*,t^*]} \to F$, $\hat{R} = R_{in} \wedge (F + W)$, *and* $W(u) = w > 0$ *for all* $u \in [s^*, t^*]$. *If* $w \geq \sup_{t \in \mathbb{R}} \{\tilde{G}(t) - (\tilde{G} * G)(t)\}$ *and* $w \geq \sup_{t \in \mathbb{R}} \{S_1(t) - (S_1 * G)(t)\}$, *then* $R_{in} \to (S_1, \tilde{G})_{[s^*,t^*]} \to R_{out}$.

Example 12 *Consider network illustrated in Figure 1, in the special case where* N_1 *represents the tandem configuration of a local GPS server with bandwidth guarantee* ϕ, *a constant propagation delay of* d_F, *and a remote GPS server with bandwidth guarantee* ϕ. *Also, suppose* N_2 *represents constant propagation delay of* d_R. *In this case,* N_1 *adaptively guarantees* $\tilde{S}_1 = \mu_\phi * \delta_{d_F}$, *and* N_2 *adaptively guarantees* δ_{d_R}. *Corollary 11 implies that if* w *is equal to the bandwidth delay product, i.e.* $w = \phi(d_F + d_R)$, *then* $R_{in} \to (\mu_\phi * \delta_{d_F}, \mu_\phi * \delta_{d_F} * \delta_{d_R})_{[s^*,t^*]} \to R_{out}$. *Furthermore, suppose that the source adapts to the feedback, such that the backlog in the throttle does not exceed* b_T^{max}. *In this case, since the backlog in network element* N_1 *is upper bounded by* $w = \phi(d_F + d_R)$, *it follows that* $R_{in} - R_{out}$ *is upper bounded by* $w + b_T^{max}$. *Hence, applying Proposition 8, it follows that the network delay is upper bounded by* $2(d_F + d_R) + b_T^{max}/\phi$.

Next, we consider the issue of how to construct network elements that provide adaptive service guarantees. We begin with a motivating proposition.

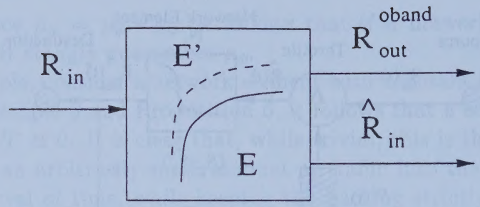


Figure 2: An Elastic Regulator

Proposition 13 (Basic Synthesis of Adaptive Service Guarantee) *Consider a network element that guarantees minimum service curve S^{min}, and suppose the arrival process is known to have the traffic envelope E. Then the network element adaptively guarantees $(S, \tilde{S})$, where $S = S^{min}$ and*

$$\tilde{S}(x) = \inf_{y:y\geq 0}\{[S^{min}(x+y) - E(y)]^+\} \quad \textit{for all } x.$$

Proposition 13 suggests a way to synthesize a network element that provides an adaptive guarantee. First, we "shape" traffic with a regulator [1][3] so that it has an envelope E. The departure process of the regulator thus has envelope E, and is then fed to a minimum service curve element with minimum service curve S^{min}. According to Proposition 13, the system then adaptively guarantee $(S^{min}, \tilde{S})$. Note that in order to maximize $\tilde{S}$, we wish to set E as small as possible. However, if E is too small, then the minimum service curve of the entire system $E * S^{min}$, will be smaller than S^{min}. This then motivates choosing $E \geq S^{min} \oslash S^{min}$. A problem with this approach is that the regulator may prevent utilization of excess bandwidth that may be offered.

3.2 Elastic Regulator

This motivates us to consider a new network element, called an "elastic regulator." which has a single arrival process, R_{in} and a departure process that is decomposed into two sub-departure processes, corresponding to a "conformant" departure process, $\hat{R}_{in}$, and an "out-of-band" departure process, R_{out}^{oband}, as illustrated in Figure 2. The elastic regulator ensures, essentially, that $\hat{R}_{in}$ has envelope E, and thus we refer to packets that are counted in $\hat{R}_{in}$ as conformant packets. In order to utilize excess bandwidth, packets may depart "out of band." Such packets are counted in the "out-of-band" departure process R_{out}^{oband}, and are called out-of-band packets.

We shall define the elastic regulator with respect to an arrival process of the form

$$R_{in}(t) = \sum_{k=1}^{\infty} L^k u(t - \tau^k),$$

where $u(x) = 1$ if $x \geq 0$ and $u(x) = 0$ if $x < 0$. In particular, packets arrive instantaneously at the instants $\tau^0 \leq \tau^1 \leq \tau^2 \cdots$, and L^k denotes the length in bits of the k^{th} arriving packet. The backlog of the elastic regulator is given by $B^r(t) = R_{in}(t) - \hat{R}_{in}(t) - R_{out}^{oband}(t)$. Packets may depart "out-of-band" at time t if $B^r(t) > 0$ or if $t = \tau^k$ for some k. We do not otherwise specify when packets may depart out of band, and in fact this is determined by a downstream network element. The elastic regulator is defined by specifying when conformant packets may leave.

Toward this end, suppose that E is a causal, sub-additive envelope. We shall in fact also assume that $E(t)$ is *left*-continuous for $t > 0$, which does not appear to be a significant loss of generality. We assume that there is a maximum packet size L_{max}, i.e. $\sup\{L^k : k \geq 0\} \leq L_{max}$.

We represent the conformant departure process as

$$\hat{R}_{in}(t) = \sum_{k=0}^{\infty} \hat{L}^k u(t - \hat{\tau}^k) ,$$

where $\hat{L}^k$ is the number of bits of the k^{th} conforming packet that departs, and we define $\hat{\tau}^0 = 0$ and $\hat{L}^0 = 0$. We also assume $0 \leq \hat{\tau}^k \leq \hat{\tau}^{k+1}$ for all k.

We define the envelope E' as follows:

$$E'(x) = \begin{cases} E(x) + L_{max} & \text{, for } x \geq 0 \\ 0 & \text{elsewhere} \end{cases} .$$

We will refer to E as the "target" envelope, since the $\hat{R}_{in}$ will not in general have envelope E, but only have envelope E'.

We describe when conforming packets may depart in terms of a sequence $\{\tau'_k\}$, defined below. We call τ'_k the eligibility time of the k^{th} conformant departure. This eligibility time is based on *previous* conformant departures from the elastic regulator and is recalculated each time *any* conformant departure takes place. The sequence $\{\tau'_k\}$ is defined as follows. Let $\tau'_1 := 0$, $\hat{\tau}^0 := 0$, and then τ'_{k+1} is computed recursively from $(\hat{\tau}^1, \hat{\tau}^2, \cdots, \hat{\tau}^k)$ at time $\hat{\tau}^k$. In particular

$$\tau'_{k+1} = \inf\{s : s \geq \hat{\tau}^k \text{ and } \min_{j:1\leq j\leq k} [\sum_{l=0}^{j-1} \hat{L}^l + E(s - \hat{\tau}^j)] \geq \sum_{l=0}^{k} \hat{L}^l\} .$$

By definition, the k^{th} conformant departure can be no earlier than τ'_k. The k^{th} confomant departure is at time τ'_k, unless $\lim_{t \to (\tau'_k)^-}(B^r_i(t)) = 0$ and there is no arrival to the elastic regulator at time τ'_k. In this case, the time of the k^{th} conformant departure is the time of the first arrival to the elastic regulator after time τ'_k.

Lemma 14 *An elastic regulator with target envelope E has the following properties:*

(i) The conformant departure process has envelope E', i.e. $\hat{R}_{in} \leq \hat{R}_{in} * E'$.

(ii) For $B^r(w) > 0$, there exists $x < w$ such that $\hat{R}_{in}(w) - \hat{R}_{in}(x) > E(w - x)$.

(iii) Given u' and any $\epsilon > 0$, there exists $w < u' + \epsilon$ and there exists $w^* \in (w - \epsilon, w + \epsilon)$ such that

$$\hat{R}_{in} * E(u') \geq \hat{R}_{in}(w) + E(u' - w) - \epsilon ,$$

$B^r(w^*) = 0$, and $\hat{R}_{in}(w) = \hat{R}_{in}(w^*)$.

Proposition 15 *Consider a system composed of an elastic regulator with target envelope E and a service curve element with minimum service curve S^{min}. The arrival process of the system directly feeds the elastic regulator. The conformant departures of the elastic regulator feed the service curve element. Finally, the departure process of the system, R_{out}, is given $R_{out} = \hat{R}_{out} + R^{oband}_{out}$, where $\hat{R}_{out}$ is the departure process of the service* *curve element and* R_{out}^{oband} *is the out-of-band departure process from the elastic regulator. If* $E \geq S^{min} \oslash S^{min}$, *there holds* $R_{in} \rightarrow (S, \tilde{S}) \rightarrow R_{out}$, *where* $S = S^{min}$ *and*

$$\tilde{S}(x) = \inf_{y:y \geq 0} \{ \left[S^{min}(x+y) - E'(y) \right]^{+} \} \quad \textit{for all } x.$$

The significance of Proposition 15 is that out of band departures from the elastic regulator allow excess bandwidth to be utilized, but do not jeopardize the benefit of the adaptive service guarantee provided by the minimum service curve element, as implied by Proposition 13. In particular, similar to [6], elastic regulators may be employed at the input of a scheduler that provides prescribed minimum service curve guarantees, as in the "SCED" scheduler in [5][10]. In this context, out-of-band packets are served when the scheduler backlog is zero. This will result in a system that allows excess bandwidth to be utilized, while providing prescribed adaptive service guarantees. Details can be found in [8]

References

[1] R. Agrawal, R. L. Cruz, C. Okino, R. Rajan, Performance bounds for flow control protocols. Technical Report ECE-98-1, ECE Dept., University of Wisconsin - Madison, May 1998.

[2] R. Agrawal and R. Rajan. A general framework for analyzing schedulers and regulators. *Proceedings of the 34th Annual Allerton Conference on Communication, Control, and Computing,* Monticello, Ill., Oct. 1996, pp. 239-248.

[3] C.-S. Chang. On deterministic traffic regulation and service guarantees: a systematic approach by filtering. *IEEE Transactions on Information Theory,* vol. 44, no. 3, May, 1998, pp. 1097-1110.

[4] R.L. Cruz. A calculus for network delay, Part I: Network Elements in Isolation, *IEEE Trans. on Information Theory,* vol. 37, no. 1, Jan. 1991, pp. 114-131.

[5] R.L. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE Journal on Selected Areas in Communications,* vol. 13, no. 6, 1995, pp. 1048-56.

[6] L. Georgiadis, R. Guérin, V. Peris, and K.N. Sivarajan. Efficient network QoS provisioning based on per node traffic shaping. *IEEE/ACM Transactions on Networking,*vol. 4, no. 4, August, 1996, pp. 482-501.

[7] J.-Y. Le Boudec. Application of network calculus to guaranteed service networks. *IEEE Transactions on Information Theory,* vol. 44, no. 3, May 1998, 1087-1096.

[8] Clayton M. Okino, A framework for performance guarantees in adaptive communication networks.. Ph.D. Thesis, Dept. of Electrical and Computer Engineering, UCSD, September 1998.

[9] A.K. Parekh and R.G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Trans. on Networking,* vol. 1, no. 3, June 1993, pp. 344-57.

[10] H. Sariowan, R. L. Cruz, and G. C. Polyzos. Scheduling for quality of service guarantees via service curves. *Proc. International Conference on Computer Communications and Networks '95 (ICCCN'95),* Sept. 1995, pp. 512-20.

Time-bounded prefix Kolmogorov complexity *
(extended abstract)

David W. Juedes
School of E.E. and C.S.
Ohio University
Athens, Ohio 45701
juedes@ohiou.edu

Jack H. Lutz
Department of Computer Science
Iowa State University
Ames, Iowa 50011
lutz@cs.iastate.edu

Abstract

In the literature, prefix Kolmogorov complexity is defined either in terms of self-delimiting Turing machines or in terms of partial recursive prefix functions. These notions of prefix Kolmogorov complexity are equivalent because, as Chaitin showed, every partial recursive prefix function can be simulated by a self-delimiting Turing machine. However, the simulation given by Chaitin's construction is not efficient, and so questions regarding the time-bounded equivalence of these notions remained unresolved. Here we closely examine these questions.

As our main result, we show that every partial recursive prefix function can be simulated with polynomial efficiency by a self-delimiting Turing machine if and only if P = NP. Thus, it is unlikely that Chaitin's construction can be used to show the polynomial-time equivalence of these notions of prefix Kolmogorov complexity.

1 Introduction

The theory of minimal programs has given valuable insight into numerous problems in disciplines ranging from combinatorics to thermodynamics. (See Li and Vitányi [12] for various examples.) The principal tool of this theory is a measure of the complexity of individual finite objects in terms of the size of programs that produce them. This tool is *Kolmogorov complexity.*

For a given finite object (string) x, the Kolmogorov complexity of x is $C(x)$, the length of the shortest program that produces x. While intuitively simple, this rough definition has an obvious shortcoming; the exact Kolmogorov complexity of an individual string is not an intrinsic property of the string itself, but is dependent upon the underlying programming model. To see this, notice that it is easy to construct a programming model where every program produces x, and hence $C(x) = 0$. As a crucial, early development, Solomonoff [13], Kolmogorov [9], and Chaitin [4] independently showed that there is an *optimal* Kolmogorov complexity measure, and hence that the Kolmogorov complexity of an individual string is indeed an intrinsic property of that string alone.

Theorem 1.1. **(Optimality Theorem)** There exists an *optimal* machine U such that for every machine M there is a constant c such that $C_U(x) \leq C_M(x) + c$ for every x.

---

*This work was supported in part by National Science Foundation Grant CCR-9157382, with matching funds from Rockwell International, Microware Systems Corporation, and Amoco Foundation.

If we restrict attention to optimal machines, then $C(x)$ becomes invariant over the choice of machine.

Theorem 1.2.(Invariance Theorem) Let U and U' be optimal machines. There exists a constant c such that $|C_U(x) - C_{U'}(x)| \leq c$ for all x.

Thus $C_U(x)$ is truly an intrinsic property of x. We often remove the dependence on U and let $C(x) = C_U(x)$ for some fixed optimal machine U.

Most straightforward variants of Kolmogorov complexity are also intrinsic properties of strings. For instance, a time-bounded version of Kolmogorov complexity is $C^t(x)$, the length of the shortest program that produces x in less than or equal to t steps. This version of Kolmogorov complexity has an *efficient* optimality theorem.

Theorem 1.3. (Efficient Optimality Theorem) There exists a machine U such that for all machines M there exists a constant c such that $C_U^{c\cdot t\log t+c}(x) \leq C_M^t(x) + c$ for all x and t.

The efficient optimality theorem serves the same purpose for time-bounded Kolmogorov complexity as the optimality theorem serves for Kolmogorov complexity; it provides a means to examine the complexity of individual strings without relying upon the architecture of the underlying machine. The efficient optimality theorem is invaluable in practice.

In this extended abstract, we examine a variant of Kolmogorov complexity for which an efficient optimality theorem is desirable — time-bounded prefix Kolmogorov complexity. Very briefly, the prefix Kolmogorov complexity of a string x is $K(x)$, the length of shortest self-delimiting program that produces x. A time-bounded version of $K(x)$ is defined in the obvious way. This variant of Kolmogorov complexity was defined independently by Levin [10], Gács [6], and Chaitin [5] and is crucial to the formulation of algorithmic probability [10, 6, 11], computational depth [3, 8], and algorithmic randomness in terms of Kolmogorov complexity [5, 14]. Similarly, time-bounded prefix Kolmogorov complexity is crucial to formulations of time-bounded algorithmic probability and computational depth.

In the literature, prefix Kolmogorov complexity is defined in terms of *self-delimiting Turing machines*, or in terms of *partial recursive prefix functions.* These versions of prefix Kolmogorov complexity are intrinsic properties of strings since they satisfy optimality theorems. Indeed, these notions of prefix Kolmogorov complexity are the *same* in the absence of time bounds. However, optimality and equivalence results for these notions of prefix Kolmogorov complexity do not easily translate to the time-bounded case. This requires some explanation.

We say that a set S is *prefix-free* if no element $x \in S$ is a proper prefix of any other element in S. A partial recursive prefix function is a partial recursive function whose domain is prefix-free. In contrast, a self-delimiting Turing machine is a Turing machine with a one-way, read-only input tape. An input x to a self-delimiting Turing machine M is *valid* if M halts with the input head reading the last bit of x. This model of computation ensures that the set of valid inputs is prefix-free.

It is natural to define prefix Kolmogorov complexity either in terms of partial recursive prefix functions or in terms of self-delimiting Turing machines. Since it is easy to construct a universal self-delimiting Turing machine, it is straightforward to prove an optimality theorem for this version of prefix Kolmogorov complexity. The proof of an optimality theorem for the version based on partial recursive prefix functions is more involved.

In an early result, Chaitin [5] shows that each partial recursive prefix function f can be simulated by a self-delimiting Turing machine M_f. We give Chaitin's complete construction below; an equivalent construction is given by Li and Vitányi [12, p. 192]. Chaitin's construction proceeds as follows. Build a self-delimiting machine M_f that reads its input one bit at a time. Let x be the portion of the input that M_f has read. If there is a proper extension of x in the domain of f, then the machine M_f reads the next bit of the input and appends it to x. If x is in the domain of f, then M_f computes $f(x)$ and halts. It is straightforward to see that the set of valid inputs of M_f is exactly the domain of f. Moreover, M_f produces the same output as f, and hence M_f computes the partial recursive function f.

Chaitin's construction may be used to construct an effective enumeration $T_1', T_2', \ldots$ of all partial recursive prefix functions. As done in [12], it is possible to use this enumeration to prove an optimality theorem for prefix Kolmogorov complexity based on partial recursive prefix functions. Notice that since every self-delimiting Turing machine computes a partial recursive prefix function, Chaitin's result also shows that the self-delimiting Turing machines compute *exactly* the partial recursive prefix functions. Moreover, Chaitin's result immediately implies that every universal self-delimiting Turing machine computes a universal partial recursive prefix function. Hence, these notions of prefix Kolmogorov complexity are the same.

While these versions of prefix Kolmogorov complexity are the same in the absence of time bounds, the relationships among these versions of prefix Kolmogorov complexity is much less clear in the presence of time bounds. Adding time bounds to prefix Kolmogorov complexity based on self-delimiting Turing machines changes very little. There exist efficient universal self-delimiting machines, and therefore proving an efficient optimality theorem for prefix Kolmogorov complexity based on self-delimiting Turing machines is straightforward. Adding time bounds to prefix Kolmogorov complexity based on partial recursive prefix functions is more problematic.

A straightforward approach to adding time bounds to prefix Kolmogorov complexity based on partial recursive prefix functions is as follows. Let M be a Turing machine that computes a partial recursive prefix function f. Then, $K_M^t(x)$ is the length of the shortest program y such that $M(x) = y$ in $\leq t$ steps. While this approach appears natural, it is problematic for several reasons. First, it is not known whether there exists a uniform means to efficiently compute every partial recursive prefix function. In particular, it is not known whether there exists an efficient universal partial recursive prefix function. Moreover, it is not known whether this version of time-bounded Kolmogorov complexity satisfies an efficient optimality theorem. Resolving these questions may be difficult.

A closer examination of the proof of the optimality theorem for prefix Kolmogorov complexity reveals the difficulty in the time-bounded case. In order to prove the optimality theorem for the version based on partial recursive prefix functions, we first convert each partial recursive prefix function to a self-delimiting Turing machine using Chaitin's construction. Since the construction given by Chaitin may dramatically increase the running time of the original machine, this approach does not give an efficient optimality theorem. With this in mind, it is natural to pose the following question.

(1) *Can Chaitin's construction be improved to show that every partial recursive prefix function can be efficiently simulated by a self-delimiting Turing machine?*

As our main result, we answer this question in the negative; every partial recursive prefix function can be simulated efficiently by a self-delimiting Turing machine if and only if P = NP. Even though the self-delimiting Turing machines compute exactly the partial recursive prefix functions, they do not do so efficiently. In the proof of our main result, we give a partial recursive prefix function that cannot be computed efficiently by a self-delimiting Turing machine unless P=NP.

Since the self-delimiting Turing machines may not be the most efficient means to compute prefix functions, it is natural to ask whether they are a reasonable basis for *time-bounded* prefix Kolmogorov complexity. The answer to this question is presently "yes" since there is no known alternative; however, the final answer to this question depends heavily on answers to the following open questions.

(2) *Does there exist an efficient universal partial recursive prefix function?*

(3) *Does there exist an efficient optimality theorem for prefix Kolmogorov complexity based on partial recursive prefix functions?*

(4) *If the answer to (3) is yes, then are the notions of prefix Kolmogorov complexity based on self-delimiting Turing machines and partial recursive prefix functions efficiently equivalent?*

This extended abstract is organized as follows. In section 2, we present preliminary notation and terminology. In section 3, we carefully review the fundamental universality and optimality theorems for time-bounded prefix Kolmogorov complexity, and we bound the difference between the two time-bounded notions of prefix Kolmogorov complexity. In section 4 we examine the connection between the complexity of computing partial recursive prefix functions with self-delimiting Turing machines and the complexity of NP. In Theorem 4.2, we prove that the partial recursive prefix functions can be computed efficiently by self-delimiting Turing machines if and only if P=NP.

2 Preliminaries

Here we follow most of the accepted conventions in the subject, such as those found in the books by Balcázar, Díaz, and Gabarró [1, 2], Hopcroft and Ullman [7], and Li and Vitányi [12]. However, we work exclusively with the alphabet $\{0,1\}$. Thus the sets $\{0,1\}^*$, $\{0,1\}^n$, and $\{0,1\}^{\leq n}$ denote the set of all strings, the set of strings of length n, and the set of strings of length $\leq n$, respectively.

We use a standard string pairing function $\langle , \rangle$ defined by $\langle x,y\rangle = bd(x)01y$, where $bd(x)$ is x with each bit doubled (e.g., $bd(101) = 110011$). Here we assume that $\langle x,y,z\rangle = \langle x, \langle y,z\rangle\rangle$, $\langle x,y,z,w\rangle = \langle x,\langle y,z,w\rangle\rangle$, etc. Note that $|\langle x,y\rangle| = 2|x| + |y| + 2$ for all $x, y \in \{0,1\}^*$.

Given two strings x and y, we write $x \sqsubseteq y$ if x is a *prefix* of y, i.e., if there exists a string z such that $xz = y$. (Note that z may be λ, the empty string.) A set X of strings is *prefix-free* if for every $x, y \in X$, $x \sqsubseteq y$ implies $x = y$.

Our primary model of computation is the Turing machine. If M is a Turing machine, we write $M(y)\downarrow$ if M on input y halts after taking finitely many steps. We write $time_M(y)$ for the number of steps that M takes on input y. We write $\mathrm{PROG}_M = \{y | M(y)\downarrow\}$ for the set of valid inputs to M. Similarly, we write $\mathrm{PROG}_M(x) = \{y | M(y) = x\}$ and $\mathrm{PROG}^t_M(x) = \{y | M(y) = x \text{ in } \leq t \text{ steps}\}$.

We are primarily interested in Turing machines whose sets of valid inputs are prefix-free. We specifically use a class of Turing machines where this condition is enforced, the *self-delimiting Turing machine.* A self-delimiting Turing machine has a single input tape, a single output tape, and $k-2$ worktapes. Each of the k-tapes is infinite and only 0's, 1's, and blanks may appear on these tapes. Each of the k tapes has a dedicated scanning head. The scanning head for the input tape is read-only and may not move left. At the start of the computation, this head is scanning a blank cell immediately to the left of the input string.

Self-delimiting Turing machines require a modified halting criterion. The computation of a self-delimiting Turing machine M on input y is a *success*, and we write $M(y)\downarrow$, if M halts after finitely many steps with the input tape head reading the last (rightmost) bit of y. Otherwise, the computation $M(y)$ is a *failure*, and we write $M(y)\uparrow$. With this modified halting criterion, it is clear that PROG_M is prefix-free for every self-delimiting Turing machine M.

If M is a standard Turing machine and PROG_M is prefix-free, then M computes a partial recursive prefix function. Notice that M need not be a self-delimiting Turing machine for this to occur. We say that M is a *universal partial recursive prefix function* if M is a partial recursive prefix function and if for every partial recursive prefix function f, there exists a string π_f such that $M(\pi_f \cdot y) = f(y)$ for every y in the domain of f.

We now precisely define prefix Kolmogorov complexity.

Definition. Let $t \in \mathbf{N}$ and let M be a standard Turing machine that computes a partial recursive prefix function.

1. The *prefix Kolmogorov complexity of $x \in \{0,1\}^*$ relative to M* is

   $$\widehat{K}_M(x) = \min\left\{|\pi| \;\middle|\; M(\pi)\downarrow \text{ and } M(\pi) = x\right\},$$

   the length of the shortest input for M that produces x. The *prefix Kolmogorov complexity of $x \in \{0,1\}^*$* is $\widehat{K}(x) = \widehat{K}_U(x)$ for some machine U that computes a universal partial recursive .

2. The *t-time-bounded (prefix) Kolmogorov complexity* of x *relative to M* is

   $$\widehat{K}^t_M(x) = \min\left\{|\pi| \;\middle|\; M(\pi) = x \text{ and } time_M(\pi) \le t\right\}.$$

   The *t-time-bounded (prefix) Kolmogorov complexity* of x is $\widehat{K}^t(x) = \widehat{K}^t_U(x)$ for some machine U that computes a universal partial recursive prefix function.

When M or U is a self-delimiting Turing machine, we write K for $\widehat{K}$.

3 Efficient Universality and Optimality

In this section we review fundamental theorems for time-bounded prefix Kolmogorov complexity and provide precise statements of the results mentioned in the introduction. We begin by asserting the existence of an efficient universal self-delimiting Turing machine.

Proposition 3.1. There exists an *efficient universal* self-delimiting Turing machine. That is, there exist a polynomial p and a self-delimiting Turing machine U that satisfy the following condition. For every self-delimiting Turing machine M, there exist a prefix

π_M and a constant c such that $U(\pi_M\pi) = M(\pi)$ and $time_U(\pi_M\pi) \leq c \cdot p(time_M(\pi))$ for every program π.

The polynomial p in Proposition 3.1 may be improved to $O(n \log n)$ using standard techniques [7]. This immediately implies an efficient optimality theorem for this version of prefix Kolmogorov complexity.

Theorem 3.2.(Efficient Optimality) There exists a self-delimiting Turing machine U such that for every self-delimiting Turing machine M, there exists a constant c such that

$$K_U^{c \cdot t \log t + c}(x) \leq K_M^t(x) + c$$

for every $x \in \{0,1\}^*$ and $t \in \mathbf{N}$. □

While the existence of an efficient universal partial recursive prefix function is open, the existence of a universal partial recursive prefix function is well known.

Proposition 3.3. (Chaitin [5]) There exists a universal partial recursive prefix function. That is, there exists a partial recursive prefix function U such that for every partial recursive prefix function f there exists a string π_f such that $U(\pi_f \cdot y) = f(y)$ for each y in the domain of f.

The existence of a universal partial recursive prefix function immediately implies optimality.

Proposition 3.4. (Optimality Theorem) There exists a partial recursive prefix function U such that for every partial recursive prefix function M there exists a constant c such that

$$\widehat{K_U}(x) \leq \widehat{K_M}(x) + c$$

for every $x \in \{0,1\}^*$.

Constructing an efficient universal partial recursive prefix function appears to be difficult. As we show in section 4, Chaitin's construction will not suffice for this unless P=NP. Moreover, it is easy to see that the set of indices of standard Turing machines that compute partial recursive prefix functions is not recursively enumerable and so a brute force approach to construct a universal machine will not work. However, this does not preclude the possibility that another construction will produce such a machine.

Here we conjecture the following.

Conjecture 3.5. There does not exist an efficient universal partial prefix function. That is, for every polynomial p, there is no universal partial recursive prefix satisfying the following; for every machine M that computes a partial recursive prefix function, there exist a string π_M and a constant c such that $U(\pi_M \cdot y) = M(y)$ for each y in the domain of M and $time_U(\pi_M \cdot y) \leq c \cdot p(time_M(y))$.

Moreover, if there does not exist an efficient universal partial recursive prefix function it seems unlikely that prefix Kolmogorov complexity could satisfy an efficient optimality theorem.

Conjecture 3.6. There does not exist an efficient optimal partial recursive prefix function. That is, for every polynomial p, there is no partial recursive prefix function U

satisfying the following; for every machine M that computes a partial prefix function there exists a constant c such that $\widehat{K}_U^{c \cdot p(t)}(x) \leq \widehat{K}_M^t(x) + c$ for each x.

As we show in section 4, both Conjecture 3.5 and 3.6 imply that $P \neq NP$, and so these conjectures are likely to be difficult to directly resolve in the affirmative. However, both conjectures may be false without resolving whether or not $P = NP$. A natural question is whether there is a complexity-theoretic hypothesis that implies either hypothesis.

If conjecture 3.6 is false, then this variant of prefix Kolmogorov complexity is indeed an intrinsic property of strings. In this case, it is natural to ask whether there is a significant difference between K and $\widehat{K}$ in the time-bounded case. Here, we provide some simple bounds on the difference between these two variants of prefix Kolmogorov complexity.

Since every self-delimiting Turing machine computes a partial recursive prefix function, it is easy to modify any universal partial recursive prefix function to efficiently simulate every self-delimiting Turing machine. Thus it is reasonable to expect that

$$\widehat{K}^{c \cdot t \log t + c}(x) \leq K^t(x) + c$$

for any reasonable choice of an efficient optimal partial recursive prefix function. (For example, the machine from Proposition 3.3 satisfies this condition.) However, it is possible that $\widehat{K}^t(x)$ is much smaller than $K^t(x)$. Here we bound the difference between K^t and $\widehat{K}^t$.

As mentioned in Li and Vitányi [12, p. 194],

$$K^t(x) \leq C^t(x) + 2 \log C^t(x).$$

This fact allows us to prove the following simple result.

Theorem 3.7. For every partial recursive prefix function M there exists a constant c such that

$$K^{c \cdot t(\log t) + c}(x) \leq \widehat{K}_M^t(x) + O(\log |x|)$$

for every $x \in \{0,1\}^*$ and $t \in \mathrm{N}$. □

4 Simulation and NP

As mentioned in the introduction, both Conjecture 3.5 and 3.6 are false if we can efficiently compute every partial recursive prefix function with a self-delimiting Turing machine. In this section we show that this avenue to resolve Conjectures 3.5 and 3.6 will not work unless P=NP. We begin by precisely formulating this hypothesis.

Hypothesis 4.1. **(The efficient simulation hypothesis)** For every partial recursive prefix function M, there exist a self-delimiting Turing machine M' and a *polynomial* p such that $M(x) = M'(x)$ and $time_{M'}(x) \leq p(time_M(x))$ for every $x \in \{0,1\}^*$.

The efficient simulation hypothesis is a statement about the complexity of computing partial recursive prefix functions by self-delimiting Turing machines. As we show in this section, the complexity of computing partial recursive prefix functions by self-delimiting Turing machines is closely tied to the complexity of NP.

Theorem 4.2. **(Main Theorem)** The efficient simulation hypothesis holds if and only if P=NP.

Proof. We begin by showing that the efficient simulation hypothesis is true if we can perform prefix searching efficiently.

Assume that P = NP, and consider the language

$$A = \left\{ \langle x, y, z \rangle \,\middle|\, \exists w \in \{0,1\}^{\leq |z|} \text{ such that } \begin{array}{ll} (i) & y \sqsubseteq w \text{ and} \\ (ii) & M_x(w) \text{ halts in } |z| \text{ steps.} \end{array} \right\}.$$

It is clear that $A \in$ NP. Since P = NP, there exists a fixed polynomial p such that $A \in \mathrm{DTIME}(p)$. Let M be a partial recursive prefix function, and let x be the index of this machine in the standard enumeration. By using x and a p-time bounded machine for A, it is reasonably straightforward to construct a self-delimiting Turing machine M' that efficiently simulates M.

To show the converse, we show that the efficient simulation hypothesis enables us to compute SAT efficiently.

To begin, construct a partial recursive prefix function M that behaves as follows.

(i) M on input $\langle x, y \rangle$ outputs 0 if x is a boolean formula with n variables and $y \in \{0,1\}^n$ is a satisfying assignment to x.

(ii) M on input $\langle x, \lambda \rangle$ outputs 1 if x is a boolean formula with no satisfying assignments.

(iii) M runs forever on any input y that does not satisfy conditions (i) or (ii) above.

By our choice of pairing function, it is clear that the set of valid inputs (programs) to M is a prefix-free set.

The machine M halts on two types of programs: programs of the form $\langle x, \lambda \rangle$, and programs of the form $\langle x, y \rangle$. M runs fastest on strings of the form $\langle x, y \rangle$. In fact, there exists a constant c such that $time_M(\langle x, y \rangle) \leq c \cdot |\langle x, \lambda \rangle|^2$, while $time_M(\langle x, \lambda \rangle) \leq c \cdot 2^{|\langle x, \lambda \rangle|}$. Since $\langle x, \lambda \rangle \sqsubseteq \langle x, y \rangle$ for every y, we can efficiently determine if $\langle x, \lambda \rangle$ is a valid program if we can efficiently simulate M on all valid programs of the form $\langle x, y \rangle$.

Assume the efficient simulation hypothesis. Fix a self-delimiting Turing machine M' and a polynomial p such that M' simulates M on all valid programs y and $time_{M'}(y) \leq p(time_M(y))$. In this case, we can use M' to solve SAT in polynomial time. Consider the algorithm given in Figure 1.

It is clear that M_{SAT} runs in polynomial time. Moreover, if x has a satisfying assignment, then there exists a y such that $M(\langle x, y \rangle)$ must accept in $\leq c \cdot |\langle x, \lambda \rangle|^2$ steps. Thus $M'(\langle x, y \rangle)$ must read past $\langle x, \lambda \rangle$ in $p(c \cdot |\langle x, \lambda \rangle|^2)$ steps. In this case, M_{SAT} accepts correctly. Otherwise, $M_{\mathrm{SAT}}(x)$ rejects correctly. It follows that M_{SAT} accepts SAT in polynomial time. Since SAT is NP-complete, it follows that P = NP. □

The complexity of NP is even more closely tied to the complexity of computing partial recursive prefix functions by self-delimiting Turing machines than Theorem 4.2 indicates. If we relax our notion of "efficient" in the efficient simulation hypothesis, then we arrive at other consequences concerning the complexity of NP. Consider the following theorem.

Theorem 4.3. Assume that there is a time-constructible function $t : \mathbf{N} \to \mathbf{N}$ such that for every partial recursive prefix function M there is a self-delimiting Turing machine

```
$M_{\mathrm{SAT}}(x)$:
begin
      if $x$ is not a boolean formula then reject;
      else
            Simulate $M'(\langle x, \lambda \rangle)$ for $p(c \cdot (|\langle x, \lambda \rangle|^2))$ steps.
            if $M'$ moves past $\langle x, \lambda \rangle$ on the input tape
               accept $x$;
            else
               reject $x$;
end;
```

Figure 1: The machine M_{SAT} from Theorem 4.2

M' such that for every $x \in \{0,1\}^*$, $M(x) = M'(x)$ and $time_{M'}(x) \leq t(time_M(x))$. Then $\mathrm{NP} \subseteq \mathrm{DTIME}(t(n^{O(1)}))$.

Proof. If we assume the hypothesis, then a straightforward modification of the construction of the machine M_{SAT} from the proof of Theorem 4.2 shows that SAT is decidable in $\mathrm{DTIME}(t(c \cdot n^2))$. Since SAT is NP-complete, it follows that $\mathrm{NP} \subseteq \mathrm{DTIME}(t(n^{O(1)}))$. $\square$

If t is a *quasipolynomial* function, i.e., $t = O(n^{(\log n)^k})$ for some k, then improving Chaitin's construction to be $t(n)$-efficient implies that

$$\mathrm{NP} \subseteq \mathrm{P}_2 = \{L | L \text{ is decidable in } \mathrm{DTIME}(f) \text{ for some quasipolynomial } f\}.$$

If t is a *subexponential* function, i.e., for every $\epsilon > 0$, $t(n) \leq 2^{n^\epsilon}$ for all but finitely many n, then improving Chaitin's construction to be $t(n)$-efficient implies that $\mathrm{NP} \subseteq \mathrm{DTIME}(2^n)$.

Since Chaitin's construction gives a simulation where prefix machines are simulated by self-delimiting machines at the expense of an exponential blowup in the running time, it is reasonable to ask if Chaitin's construction can be improved even slightly. Theorem 4.3 shows that even a slight improvement separates NP from E_2.

5 Conclusion

It is unlikely that self-delimiting Turing machines can efficiently compute all partial recursive prefix functions since this implies that $\mathrm{P} = \mathrm{NP}$. Unfortunately, this leaves several important questions unresolved even if $\mathrm{P} \neq \mathrm{NP}$.

(i) Is there an efficient universal partial recursive prefix function?

(ii) Is there an efficient optimal theorem for prefix Kolmogorov complexity based on partial recursive prefix functions?

(iii) Are the two notions of prefix Kolmogorov complexity efficiently equivalent?

Prefix Kolmogorov complexity based on self-delimiting Turing machines has an efficient optimality theorem and therefore is an intrinsic property of individual strings.

Even though they may not efficiently compute all partial recursive prefix functions, self-delimiting Turing machines appear to be the only reasonable basis for time-bounded prefix Kolmogorov complexity in the absence of affirmative answers to questions (i) and (ii).

Acknowledgments The authors thank Harry Buhrman and Giora Slutzki for helpful discussions.

References

[1] J. L. Balcázar, J. Díaz, and J. Gabarró. *Structural Complexity I.* Springer-Verlag, Berlin, 1988.

[2] J. L. Balcázar, J. Díaz, and J. Gabarró. *Structural Complexity II.* Springer-Verlag, Berlin, 1990.

[3] C. H. Bennett. Logical depth and physical complexity. In R. Herken, editor, *The Universal Turing Machine: A Half-Century Survey*, pages 227–257. Oxford University Press, Oxford, 1988.

[4] G. J. Chaitin. On the length of programs for computing finite binary sequences: statistical considerations. *Journal of the ACM*, 16:145–159, 1969.

[5] G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the Association for Computing Machinery*, 22:329–340, 1975.

[6] P. Gács. On the symmetry of algorithmic information. *Soviet Mathematics Doklady*, 15:1477–1480, 1974.

[7] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading, Mass., 1979.

[8] D. W. Juedes, J. I. Lathrop, and J. H. Lutz. Computational depth and reducibility. *Theoretical Computer Science*, 132:37–70, 1994.

[9] A. N. Kolmogorov. Three approaches to the quantitative definition of 'information'. *Problems of Information Transmission*, 1:1–7, 1965.

[10] L. A. Levin. Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problems of Information Transmission*, 10:206–210, 1974.

[11] L. A. Levin. Randomness conservation inequalities; information and independence in mathematical theories. *Information and Control*, 61:15–37, 1984.

[12] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications, Second Edition.* Springer–Verlag, New York, 1997.

[13] R. J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7:1–22, 224–254, 1964.

[14] V. V. V'jugin. The algebra of invariant properties of finite sequences. *Problems of Information Transmission*, 18:147–161, 1982.

Multisignature Schemes for Various Group Structures

Hiroshi Doi
Department of Mathematics, Faculty of Science, Okayama University
3-1-1 Tsushima-naka, Okayama, 700-8530 Japan
hdoi@math.okayama-u.ac.jp

Eiji Okamoto
JAIST (Japan Advanced Institute of Science and Technology)
1-1 Asahidai, Tatsunokuchi, Nomi, Ishikawa, 923-1292 Japan
okamoto@jaist.ac.jp

Masahiro Mambo
Education Center for Information Processing and
Graduate School of Information Sciences, Tohoku University
Kawauchi, Aoba, Sendai, 980-8576 Japan
mambo@ecip.tohoku.ac.jp

Abstract

A multisignature scheme is a type of digital signature scheme in which more than one user sign a single message. In many cases, each signer is in a different position in a signing group and has a different responsibility. The signing order often reflects these differences. For example, a user U_4 supervising the work of a user group should sign a document after all members U_1, U_2, U_3 of the user group has signed. In such a case, any multisignature not obeying this signing order should become invalid. In this paper, we propose two multisignature schemes suitable for various group structures.

First, we propose a multisignature scheme whose underlying signature is composed of two values, such as ElGamal, Schnorr and DSA. Although the signature length expands as the number of signers increases, the security of the proposed scheme is considered to be enough high compared with the underlying signature scheme.

Secondly, we propose a common modular RSA multisignature scheme. The security level of this signature scheme is a little weaker than the first multisignature scheme. But it can avoid signature expansion and the signatures for different signing orders can be verified by only one set of verification keys. This scheme seems to fit well with various cryptographic techniques, like the threshold technique. But we point out that direct application of the threshold technique is actually not secure.

1 Introduction

A multisignature scheme is a type of digital signature scheme in which more than one user sign a single message. Multisignature schemes are constructed using a public key cryptosystem, for example RSA [10]. Multisignature schemes based on RSA can be divided into two classes. One is a scheme in which all signers use common modulus such as [1] or [3], the other is a scheme in which each signer uses different modulus such as [9]. Besides RSA, there is a multisignature scheme [13] based on ElGamal [6].

Generally speaking, each signer of a multisignature scheme is in a different position in a signing group and has a different responsibility. The signing order often reflects these differences. For example, a user U_4 supervising the work of a user group G should sign a document after all members U_1, U_2, U_3 of G has signed. In such a case, if U_4 signs before other group members, a generated multisignature should become invalid. Nonetheless, the signing order does not have meaning in all cases. It could happen that the signing order in G has no importance. Then U_1, U_2, U_3 should sign in an arbitrary order and U_4 should sign afterwards.

The signing order of multisignature schemes has not been paid much attention so far except [4, 5, 7, 11, 14]. In these schemes, [4, 5] is suitable for various group structures. But its method can be applied only to RSA and its security is not evaluated.

Therefore, we propose two multisignature schemes suitable for various group structures whose security is evaluated under the following conditions.

$C1$: **Versatility** Proposed scheme is suitable for various group structures composed with parallel and serial order.

$C2$: **Non reusability** Intermediate signature for one group structure cannot be reused for another group structure.

$C3$: **Validity of multisignature** Multisignatures are accepted only when they are generated by correct signers in a correct order.

$C4$: **Validity of intermediate signature** Before signing a document, any signer can verify signatures generated by the former group of signers.

2 Notation

$A, B, C, \cdots$ are signers. M is a message to sign. All signer U publish a pair of identities (U_1, U_2). U_1 and U_2 are integers modulo some big number. In this ring, arithmetical operation can be defined for all units. A signature which is composed of two values such as ElGamal, Schnorr[12] and DSA[2] is called ElGamal-type signature in this paper. Similarly, a multisignature derived from the ElGamal-type signature is called ElGamal-type multisignature. An ElGamal-type signature of U is expressed by (u_1, u_2).

Conditions of Identities: A user U has a pair of (pseudo) identities (U_1, U_2) which depends on each other. This means that no one can simultaneously select both of U_1 and U_2 of his own choice. We can use a hash function for computing U_2 from U_1. Or we can use ElGamal-type signatures for an appropriate message M. U_1 and U_2 are taken to be sufficiently large so that no collision is expected among (U_1, U_2).

We assume the existence of (target) collision-resistant hash functions $h(x)$. When x is a concatenation of two inputs x_1 and x_2, $h(x)$ is written by $h(x_1, x_2)$. h is a published function.

3 Structured Group ID

There are two kinds of basic group structures, parallel and serial. All of group structures are constructed as a combination of these basic structures. We introduce a notion of a group identity GID derived from the group structure.

3.1 Serial Group Structure

In the serial group structure, different order of signers cause different GID. We construct multisignature schemes using such a GID. Consider the most simple case such that a user B must sign a message after A. If B has signed before A, its GID and the corresponding multisignature should become totally different from GID and multisignature for the expected order A, B. By such a construction multisignatures for different order can be distinguished.

We describe how to make GID for serial group structure. All computation for GID is done modulo some enough large integer. A has (A_1, A_2) and B has (B_1, B_2) as ID. If the order is A, B, then A makes his partial GID $I_A = (1 + A_1)A_2$ and B makes his GID $I_B = (I_A + B_1)B_2$ using I_A. So, the GID for the order A, B is $((1 + A_1)A_2 + B_1)B_2$. On the other side, if the order is B, A, the GID is $((1 + B_1)B_2 + A_1)A_2$.

Proposition 1 *If $|A_1|$, $|A_2|$, $|B_1|$, $|B_2|$ are large enough, the probability of selecting (A_1, A_2) and (B_1, B_2) which work both for the order of A, B and the order of B, A is negligibly small.*

Proof: By solving $((1+A_1)A_2+B_1)B_2-((1+B_1)B_2+A_1)A_2 = 0$, we have $A_1A_2(B_2-1)-B_1B_2(A_2-1) = 0$. At first, A_1 and A_2 are fixed by a pair of mutually related values. Then $B_2(C-DB_1) = C$ is derived, where $C = A_1A_2$ and $D = A_2-1$. Recall that B_1 and B_2 are also mutually related. Assuming the random behavior of one identity in mutual relation and $N = |U_1| \simeq |U_2|$, the probability of computing a correct pair of (B_1, B_2) is roughly estimated to be less than $\frac{1}{2^{N/2}}$ by the birthday paradox. Then the probability of selecting a tuple (A_1, A_2, B_1, B_2) valid for both order is roughly less than $\frac{1}{2^{3N/2}}$. Because N is taken to be large enough, the probability can be negligibly small. The observation described above can be extended into a more general case. $((S+A_1)A_2+B_1)B_2-((S+B_1)B_2+A_1)A_2 = 0$, where S is the GID generated by former signers. ∎

3.2 Parallel Group Structure

In the parallel group structure, the difference in signing order among signers has no importance. Moreover, each signer can make his GID independently and simultaneously. After each signer makes his GID, someone collects them and constructs a GID for parallel group structure. In the most simple case, there is a situation that users A, B sign the same message at the same time. In such a case, the order of signers has obviously no meaning.

Now we describe how to make GID for parallel group structure. A has (A_1, A_2) and B has (B_1, B_2) as own ID. The partial GID made by A and B for the group is defined by $(1 + A_1)A_2$ and $(1 + B_1)B_2$, respectively. Then a third person or one of A and B collecting A and B's GID computes $(1 + A_1)A_2 + (1 + B_1)B_2$ as the GID for a parallel group A, B.

For discussing a more general case, we set S as the partial GID constructed by former signers. A and B make his GID simultaneously after the generation of S. Then a combined GID is described as $(S + A_1)A_2 + (S + B_1)B_2$.

3.3 Various Group Structure

Serial and parallel group structures can be combined as shown in figure 1. In figure 1, the signing order inside each group is not particularly important. But different orders of groups result in different signatures.

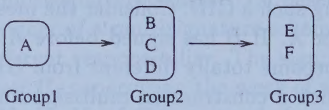


Figure 1: Example 1

First, A of group 1 creates his GID $S_1 = (1 + A_1)A_2$. Secondly, each user of group 2 gets S_1 and creates his GID in parallel. User's GID are collected and the combined GID S_2 of group 2 is computed by $S_2 = (S_1 + B_1)B_2 + (S_1 + C_1)C_2 + (S_1 + D_1)D_2$.

Repeating the same procedure as in group2, the GID for group 3 is obtained by

$$\sum_{V\in\{E,F\}} (\sum_{U\in\{B,C,D\}} (((1 + A_1)A_2) + U_1)U_2 + V_1)V_2).$$

On the other hand, there is a group structure which has multiple serial structures as shown in figure 2.

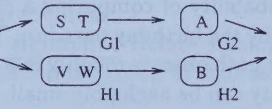


Figure 2: Plural Serial Structure

A signs a message M after S, T in group $G1$. Independently and simultaneously, B signs M after V, W of group $H1$. A collecting person computes final GID of the above example:

$$(((1 + S_1)S_2 + (1 + T_1)T_2) + A_1)A_2 + (((1 + V_1)V_2 + (1 + W_1)W_2) + B_1)B_2.$$

4 ElGamal-Type Multisignature

4.1 Proposed Scheme

ElGamal-type multisignature scheme can be written by induction.

Signature of the first group: A user U in the first group knows his position in the group structure. U uses published user IDs to make his GID, say I_U, for a specified group. Next, U calculates $h(I_U, M)$ and generates a signature for $h(I_U, M)$. U's signature is (u_1, u_2) and U sends it to the next group.

Signature of intermediate groups: When user V in the intermediate group gets M and former signatures, he verifies the validity of the former signatures at first. Because V knows published IDs, he can check the validity of the former signatures. Next V makes his GID as the same manner in the first group. The only difference is to use signatures of the former signers in place of IDs of the former signers. This means that V's GID can only be generated using former signatures. He generates his signature (v_1, v_2) as the same manner of the first group.

Verification: A verifier gets M and a multisignature for M signed by all users in the groups. He uses them and published ID to check each component signature. The checking method is the same as a verification method of users in the intermediate groups.

In the above scheme, $C1$ and $C4$ are satisfied.

About $C2$, all signers sign the hashed value which depends on GID. If the group is different, GID is different by proposition 1, therefore the hashed value is different. So it is difficult to reuse the intermediate signature for another group.

About $C3$, the intermediate signature depends on the former signatures. When the signer did not observe the signing order, it will be hard for the verifier to accept his signature.

At present, no counter-example has been found for $C2$ and $C3$. As a negative property, the length of signature increases as many users participate in the group.

Even if underlying signature is composed of one value, it is easy to generate second value by hash function. So, ElGamal-Type multisignature scheme can be applied to the signature scheme as RSA.

4.2 Example

The ElGamal-type multisignature is exemplified for a group structure shown in figure 1.

Signature of A: A creates his GID

$$I_A = (((A_2 + A_1A_2)(B_2 + C_2 + D_2) + B_1B_2 + C_1C_2 + D_1D_2)(E_2 + F_2) + E_1E_2 + F_1F_2).$$

Next, A calculates $h(I_A, M)$ and creates a signature (a_1, a_2) for $h(I_A, M)$. Then he sends it to the next group.

Signature of B, C, D: Each B, C, D gets $M, (a_1, a_2)$. First, B checks the validity of (a_1, a_2). Then B uses published IDs and (a_1, a_2) to make his GID

$$I_B = (((a_2 + a_1a_2)(B_2 + C_2 + D_2) + B_1B_2 + C_1C_2 + D_1D_2)(E_2 + F_2) + E_1E_2 + F_1F_2).$$

Next, B calculates $h(I_B, M)$ and generates a signature (b_1, b_2) for $h(I_B, M)$. C and D do the same. Their GIDs become the same, $I_B = I_C = I_D$. B, C, D can generate their signatures independently and simultaneously. B, C, D send the former signatures and their signatures to the next group.

Signature of E, F: Each of E, F gets M and the former signatures, (a_1, a_2), $\cdots$ (d_1, d_2). At first, E checks the validity of (a_1, a_2), $\cdots$, (d_1, d_2). Then E makes his GID

$$I_E = (((a_2 + a_1a_2)(b_2 + c_2 + d_2) + b_1b_2 + c_1c_2 + d_1d_2)(E_2 + F_2) + E_1E_2 + F_1F_2).$$

Next, E calculates $h(I_E, M)$ and generates a signature (e_1, e_2) for $h(I_E, M)$. F does the same. E, F send the former signatures and their signatures to the next group.

Verification: A verifier gets M and (a_1, a_2), $\cdots$, (f_1, f_2). He checks them using published IDs.

5 Common Modular RSA Multisignature

There are a lot of multisignature schemes based on RSA. In the scheme described in [9], an intermediate signer can verify that the validity of former signatures and the length of signature is almost equal to the original signature scheme. In this scheme, modulus n is different for each signer.

On the other hand, there are many kinds of RSA multisignature schemes [1, 3] which use a common modulus n. Their methods use arithmetical operations on the exponential part. In the common modular RSA schemes, the validity of partial signature cannot be verified until all signers finish signing a message. Furthermore, the existence of trusted center must be assumed in many schemes. Even so, common modular RSA is a very interesting research topic.

As observed in Sect.3, only addition and multiplication modulo an integer are required for calculating GID. Fortunately, these arithmetic can be applicable in the exponent part of common modular RSA schemes. There is a paper [4, 5] in this direction, and we improve this method.

A user U publishes his mutually related ID (U_1, U_2). He has a secret key u_1 which is independent on his ID. We assume that there is a trusted center which knows secret keys u_1's of all users U's. The center publishes a common modulus n of RSA scheme.

5.1 Proposed Scheme

A common Modular RSA multisignature scheme can be written by induction. In the following calculation, if we use modular arithmetic, the modulus is n in default.

Preparation: A trusted center distributes a modulus n and a verification key v_1, v_2 for an entire group of multisignature. The center must calculate a verification key to satisfy the congruence in the verification step.

Signature of first group: Each user U in the first group, say F, knows the group structure. At first, U uses the published ID to make his GID, say I_U. Next, U calculates another secret key $u_2 = h(I_U, u_1)$ dependent on GID and u_1, and generates his signature for $\overline{M} = h(I_U, M)$. U's signature is $(\overline{M} \cdot \overline{M}^{u_1})^{u_2}$. Before they send the signature to the next group, someone must multiply their signatures to create

$$S \equiv \overline{M}^{\sum_{U \in F} u_2 + u_1 u_2}.$$

Signature of intermediate groups: User V in the intermediate group, say G, gets M and the former signature S. He makes his GID, say $I_V (= I_U)$, as the same manner in the first group and constructs $v_2 = h(I_V, v_1)$ and $\overline{M} = h(I_V, M)$. Then V generates his signature as $(S \cdot \overline{M}^{v_1})^{v_2}$. Before they send it to the next group, someone must multiply their signatures to create new S as follows.

$$S \equiv S^{\sum_{V \in G} v_2} \cdot \overline{M}^{\sum_{V \in G} v_1 v_2}$$

Verification: A verifier gets M and a multisignature S. Using published n, v_1 and v_2 he checks the following congruence

$$(S \cdot \overline{M}^{v_1})^{v_2} \equiv \overline{M}.$$

The left side member of the above congruence is $\overline{M}^{(x+v_1)v_2}$ where x is decided by the group structure and all users secret keys. So the trusted center can get appropriate v_1, v_2. If the above congruence is satisfied, the multisignature is accepted.

Note that the common modulus n must be changed for different group structure [8]. So the center must distribute n, v_1, v_2 when new group structure is constructed.

The above scheme has the following features.

1. This scheme satisfies $C1$.
2. The evaluation for $C2$ is as same as ElGamal-Type multisignature. Furthermore, modulo n is different if the group structure is different.
3. There is no method to verify the intermediate signature. So this scheme does not satisfy $C4$.
4. A signer in the same group can use same modulus n and verification key v_1, v_2.
5. The length of signatures does not increase for any group.

In [4, 5], user's secret keys (u_1, u_2) are simple. In [4], $u_1 = 1$ and u_2 is arbitrary. On the other hand, both secret keys are arbitrary in [5]. So there are some attacks related to $C3$. About $C3$, we can prove the following propositions for the proposed scheme.

Proposition 2 *Even by the collusion of all but one signer forging a valid multisignature is as difficult as forging a RSA signature.*

Proof: Define by $\mathtt{RSAsig}(M, e, n)$ a function that on input $M \in \mathbf{Z}_n^*$, $e \in \mathbf{Z}_{\phi(n)}^*$, $n \in \mathbf{N}_{>1}$, outputs S such that $S^e \equiv M \bmod n$ if such a S exists.

Define by $\mathtt{PROPsig}(M, IGS, u_{i1}, u_{i2}, v_1, v_2, n)$ a function that on input $M \in \mathbf{Z}_n^*$, information for group structure IGS, $(u_{i1} \in \mathbf{Z}_{\phi(n)}, u_{i2}(= g(u_{i1})) \in \mathbf{Z}_{\phi(n)})$ for all but one signer, where g is some function, $(v_1 \in \mathbf{Z}_{\phi(n)}, v_2 \in \mathbf{Z}_{\phi(n)}^*)$, $n \in \mathbf{N}_{>1}$, outputs S such that $(S \cdot M^{v_1})^{v_2} \equiv M \bmod n$ if such a S exists. Note that $S \equiv M^{x+(y+w_1)w_2 z} \bmod n$, where x, y, z are computed from all (u_{i1}, u_{i2}) but $(w_1, w_2(= g(w_1))$ based on IGS.

1. $\mathtt{RSAsig} \le_{1-tt} \mathtt{PROPsig}$:
   $\mathtt{RSAsig}(M, e, n) = \mathtt{PROPsig}(M, IGS, u_{i1}, u_{i2}, e, e, n) M^e \bmod n$, where u_{i1}, u_{i2} are randomly selected and IGS is information on arbitrary group structure. Since there exists $(w_1, w_2)'s$ satisfying $(x + (y + w_1)w_2 z + v_1)v_2 \equiv 1 \bmod \phi(n)$ in almost all the cases, `PROPsig` returns an answer.
2. $\mathtt{PROPsig} \le_{1-tt} \mathtt{RSAsig}$:
   $\mathtt{PROPsig}(M, IGS, u_{i1}, u_{i2}, v_1, v_2, n) = \mathtt{RSAsig}(M, v_2, n)/M^{v_1} \bmod n$.

Thus $\mathtt{PROPsig} \equiv_{1-tt} \mathtt{RSAsig}$. ∎

Assume that all users except only one user W collude. If W get former signature $S_1 = M^x$ for inappropriate x, he cannot know the value x. So, W may sign S_1 and pass his signature S_W to the next group. Can the appropriate multisignature be calculated from S_W? The following proposition says some kind of calculation cannot be executed.

Proposition 3 *If it is infeasible to compute $\phi(n)$ from n and (u_{i1}, u_{i2})'s of all but one signer W, then it is infeasible to compute exponents $x + (y + w_1)w_2 z \bmod \phi(n)$ of valid multisignatures for more than three group structures.*

Proof: If more than three valid exponents can be computed, we have the following congruences modulo $\phi(n)$ by simple calculation. Unknown variables are w_1, w_2 and $\phi(n)$.

$$\begin{aligned}
((x^{(1)} + y^{(1)}w_2 + z^{(1)}w_1 w_2) + v_1)v_2 &\equiv 1, \\
((x^{(2)} + y^{(2)}w_2 + z^{(2)}w_1 w_2) + v_1)v_2 &\equiv 1, \\
((x^{(3)} + y^{(3)}w_2 + z^{(3)}w_1 w_2) + v_1)v_2 &\equiv 1,
\end{aligned}$$

where $x^{(i)}, y^{(i)}, z^{(i)}$ are computed from (u_{i1}, u_{i2})'s of all signers except W. After simple calculation, we can remove the term w_2 and w_1w_2 from these congruences. Finally, colluding signers can get some value X congruent to 0. If the value X is not equal to 0, the multiple of $\phi(n)$ can be computed. This contradict the assumption on the infeasibility of computing $\phi(n)$. ∎

Concerning $C3$, there may be a method of calculating S' and some function f s.t. $(f((S'\overline{M}^{c_1})^{c_2}) \cdot \overline{M}^{v_1})^{v_2} \equiv \overline{M}$. We cannot prove the security on such an attack at present.

5.2 Example

The multisignature for the figure 1 is generated as follows. Note that a modulus n and a pair of verification keys v_1, v_2 are distributed by the trusted center.

Signature of A: A uses published IDs to make his GID as follows.

$$I_A = (((A_2 + A_1A_2)(B_2 + C_2 + D_2) + B_1B_2 + C_1C_2 + D_1D_2)(E_2 + F_2) + E_1E_2 + F_1F_2)$$

Next, A calculates $a_2 = h(I_A, a_1)$, $\overline{M} = h(I_A, M)$ and generates his signature $S_1 = S_A \equiv (\overline{M} \cdot \overline{M}^{a_1})^{a_2}$.

Signature of B, C, D: Each B, C, D gets M, S_1. B uses published IDs to make his GID $I_B (= I_A)$ as computed by A. Next, B calculates $b_2 = h(I_B, b_1), \overline{M}$ and generates a signature for $\overline{M}$. His signature is $S_B \equiv (S_1 \cdot \overline{M}^{b_1})^{b_2}$. Each C and D generates his signature S_C and S_D, respectively. Before their signatures are passed to the next group, someone multiplies each signature. The following $S_2 = S_BS_CS_D$ is the intermediate signature for group 1 and 2.

$$S_2 \equiv \overline{M}^{((a_2+a_1a_2)(b_2+c_2+d_2)+b_1b_2+c_1c_2+d_1d_2)}$$

Signature of E, F: Each E, F gets M, S_2. E makes I_E and e_2 as computed by A, B, C and D. His signature is $S_E \equiv (S_2 \cdot \overline{M}^{e_1})^{e_2}$. F generates his signature S_F, either. The multisignature $S_3 = S_ES_F$ for group 1, 2 and 3 is

$$\begin{aligned} S_3 &\equiv \overline{M}^x, \quad where \\ x &= (((a_2 + a_1a_2)(b_2 + c_2 + d_2) + b_1b_2 + c_1c_2 + d_1d_2)(e_2 + f_2) + e_1e_2 + f_1f_2). \end{aligned}$$

Verification: A verification key must be distributed from the trusted center when n are distributed. The trusted center knows all users secret keys, hash function and the group structure. So he can create the second secret keys and create v_1, v_2 satisfying the following congruence modulo $\phi(n)$.

$$(((a_2 + a_1a_2)(b_2 + c_2 + d_2) + b_1b_2 + c_1c_2 + d_1d_2)(e_2 + f_2) + e_1e_2 + f_1f_2 + v_1)v_2 \equiv 1.$$

A verifier gets M and multisignature S_3. He checks the validity of $(S_3 \cdot \overline{M}^{v_1})^{v_2} \equiv \overline{M}$.

5.3 Insecurity of Applying Threshold Scheme

Can the threshold scheme in [3] be applied to our scheme? As in the [3], the modulus n must be equal to pq where $p = 2p' + 1$, $q = 2q' + 1$ and p, q, p', q' are primes. Let d be the secret key and k be the threshold, respectively. In [3], Each user has his secret key distributed from the trustd center and any k users can create M^d for an arbitrary M. In this procedure, their secret keys are distributed only once at the preliminary phase.

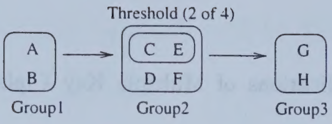


Figure 3: Group Structure with Threshold

In figure 3, the shreshold k is equal to 2. First, users in group 1 sign M. After group 1, users belonging to group 2 sign by the 2 out-of 4 threshold scheme. Finally, all users of group 3 sign. In this case, the trusted center also generates secret numbers s_1, s_2 for group 2 and verification keys v_1, v_2 satisfying

$$(((((a_2+b_2)+(a_1a_2+b_1b_2))s_2+s_1s_2)(g_2+h_2)+g_1g_2+h_1h_2)+v_1)v_2 \equiv 1 \; mod \; \phi(n),$$

where n is the modulus generated by the trusted center for the threshold scheme described above. The center distributes the partial secret for s_1 and s_2 to each C, D, E, F. a_1, a_2, b_1, b_2, g_1, g_2 and h_1, h_2 are secret keys of A, B, G and H, respectively.

All users in group 2 received S_1 as the signature of group 1. Each of two users generate $\overline{M}^{s_1}$ using threshold scheme. Next, the same two users generate $(S_1\overline{M}^{s_1})^{s_2}$ using threshold scheme again. Another procedure and verification are done as in the non-threshold version.

But this scheme is weak against the collusion. If $k+1$ users in the threshold group collude, arbitrary k users of the colluding users can generate the partial secret s_1 mod $\phi(n)$. If these partial secret is not equal, the colluding users can get some multiples of $\phi(n)$. As well known, $\phi(n)$ may be revealed by calculation of G.C.D. of these values. Furthermore, if ((multiple of $\phi(n)$), v_2) $= 1$, an appropriate S satisfing $(S\overline{M}^{v_1})^{v_2}$ can be calcureted by the users of the threshold group. Therefore, direct application of the threshold scheme should be avoided.

6 Conclusion

In this paper, we have discussed the order of signature of serial and parallel group structure and introduced the group identity uniquely specified by the group structure. We also constructed multisignature schemes for various group structures.

The first scheme is based on ElGamal-type signature scheme which satisfies almost all conditions described in section 1. As a drawback the total length of signature is very large compared with the original signature scheme.

The second scheme is based on a common modular RSA signature scheme which satisfies $C1, C2$ described in section 1. Although it needs several assumption such as a trusted center, the length of signature does not increase and several applications can be considered for it. But it turns out that when threshold scheme is directly applied to the proposed scheme, an obtained scheme is insecure.

7 Acknowledgments

The authors are grateful to Mike Burmester, Yvo G. Desmedt, Xun Yi, Yuko Yoshifuji and Takashi Tasaka for useful discussion and advice.

References

[1] C. Boyd, "Some Applications of Multiple Key Ciphers," *Proc. of Eurocrypt'88*, pp. 455-467, 1988.

[2] "Debating Encryption Standards", *Communications of ACM*, Vol. 35, No. 7, pp. 32-54, 1992.

[3] Y. Dsemedt and Y. Frankel, "Shared generation of authenticators and signatures," *Proc. of Crypto '91*, pp. 457-469, 1991.

[4] H. Doi, E. Okamoto, M. Mambo and T. Uyematsu, "Multisignature Scheme with Specified Order." in *1994 Symposium on Cryptography and Information Security*, SCIS94-2A, Japan, January 27-29, 1994.

[5] H. Doi, "The application of RSA for multisignature schemes and substantial number of cryptographic keys," *Master Thesis*, JAIST, 1994.

[6] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE Trans. on Inform. Theory*, Vol. IT-31, No. 4, pp. 469-472, 1985.

[7] M. Mambo and E. Okamoto, "A Note on Multisignatures," in *1994 Symposium on Cryptography and Information Security*, SCIS94-2B, Japan, January 27-29, 1994.

[8] J. H. Moore, "Protocol Failures in Cryptosystems," *Proc. of the IEEE*, Vol. 76, No. 5, pp. 594-602, May 1988.

[9] T. Okamoto, "A Digital Multisignature Scheme Using Bijective Public-Key Cryptosystems," *ACM Trans. on Computer Systems*, Vol. 6, No. 8, pp. 432-441, Nov. 1988.

[10] R. L. Rivest, A. Shamir, L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Commun. of the ACM*, Vol. 21, No. 2, pp. 120-126, 1978.

[11] T. Saito, "A Multisignature Scheme Enabling a Specified Signer's Order," in *1997 Symposium on Cryptography and Information Security*, SCIS97-33A, Japan, January 29-February 1, 1997.

[12] C. P. Schnorr, "Efficient identification and signatures for smart cards," *Proc. of Crypto'89*, pp. 239-252, 1990.

[13] A. Shimbo, "Multisignature Schemes Based on the ElGamal Scheme," in *1994 Symposium on Cryptography and Information Security*, SCIS94-2C, Japan, January 27-29, 1994.

[14] Y. Yoshifuji, M. Burmester, H. Doi, M. Mambo, M. Tada and E. Okamoto, "An Order-Specified Multisignature Scheme based on a Modified ElGamal Signature Scheme," in *Computer Security Symposium '98*, CSS'98-5A-2, Japan, 1998.

QUANTUM INFORMATION PROCESSING

Gilles Brassard
Department IRO
University of Montreal
Montreal, Quebec H3C 3J7
Canada
brassard@IRO.UMontreal.CA

Quantum mechanics teaches us that things do not behave at the scale of elementary particles the way we are used to in our macroscopic experience. The fact that classical information theory is firmly rooted in classical physics has prevented us from tapping the full potential of physical reality for information processing purposes. Quantum information is very different from its everyday classical counterpart: it cannot be read without disturbance, it cannot be copied or broadcast at all, but it can exist in superposition of classical states. Classical and quantum information can perform together feats that neither could achieve alone, such as quantum cryptography, quantum computing and quantum teleportation.

Quantum cryptography exploits the unavoidable disturbance caused by any attempt at measuring quantum information to implement a cryptographic system that allows two people to communicate in absolute secrecy. Quantum computers take advantage of the superposition principle to allow exponentially many computations to be performed simultaneously in a single piece of hardware. Quantum teleportation uses nonlocal effects of quantum mechanics to allow the transmission of quantum information with no need for a quantum channel at the time transmission actually occurs, and it can be used for the faithful transmission of quantum information on an unreliable channel.

Planning and Control for Planar Batting and Hopping

Craig K. Black   Kevin M. Lynch

Laboratory for Intelligent Mechanical Systems
Northwestern University
Evanston, Illinois 60208 USA

Abstract

Manipulation by batting and locomotion by hopping share several common features. We look at planning and control methods for these types of dynamic underactuated manipulation using predictive model optimization. This generates optimal control sequences for reaching a desired goal state. Results are given for disk batting simulations. Current work is focused on experimental implementation.

1. Overview

Manipulation by batting and locomotion by hopping share several common features. Hopping can be viewed as a type of "self manipulation," and the relationship between batting and hopping has been studied in depth by Koditschek and colleagues [1,2,5,12]; see also (M'Closkey and Burdick [8]; Spong [14]; Vakakis *et al.* [17]). Batting and hopping are both examples of intermittent dynamical systems. Most previous work on controlling such systems has focused on empirically derived control laws for regulating cyclic batting to a fixed point [1,2,5,12] and (Schaal and Atkeson [13]; Zumel and Erdmann [20]) and hopping with a fixed height and forward speed (Raibert [10]). These controllers have been studied experimentally [1,2,10,12] and (Hodgins and Raibert [4]) and their local and global stability properties have been analyzed [5,8,11,17,20].

Our work is aimed at deriving control laws to transfer the system from a given initial state (of the batted object or hopper) to a desired goal state. In the case of batting, a desired goal state might be to bat a ball into a hoop or into another robot's workspace; in the case of a hopper, the goal could be to land on a stepping stone or to perform a flip. Because external control forces can only be applied during the brief contact phase, generally more than one bat or one hop will be required to carry the system to the goal state, requiring the control law to have a form of "look ahead" embedded in it. Interesting problems include determining the number of impact events (bats or hops) necessary to reach the goal state; characterizing the geometry of the system state space accessible from the initial state; and deriving a suitable control law. Ideally the control law would find a control sequence to take the system to the goal state when it is reachable, and reduce to a (nearly) globally convergent control law when the goal state is reachable from successive impact events (as with the mirror law for planar disk juggling [1,2]).

We have built a one joint robot which bats planar parts floating on an air table in a gravitational field (Figure 1). 60 Hz vision feedback is provided by an overhead camera. This system is quite similar to the planar juggler of Bühler and Koditschek [1,2]. The goal is to derive a controller to bat parts from a given initial state to a desired state. This paper describes our preliminary work toward this goal. Our controller uses a model of planar impact dynamics in a nonlinear optimization to find sequences of bats that, in simulation, transfer parts to specified goal states. Our current work is aimed at validating the approach on our experimental setup.

Our work is motivated in part by Brown and Zeglin's [19] bow leg planar hopping robot. The robot consists of a low mass fi-

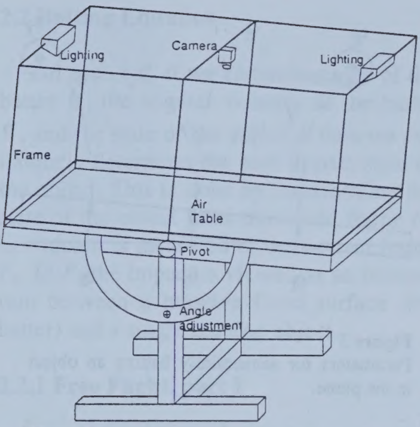


Figure 1
Our experimental testbed, Flatland. Planar objects float on the air table, which is tiltable to vary the effects of gravity in the plane. Modular 1 and 2 DOF robots are easily mounted to the side framing. The overhead camera tracks the object position and orientation.

berglass leg and a relatively high mass robot body, so that the leg can be positioned during flight without affecting the body's motion. The bow leg can be cocked by retracting a string attached to the end of the bow, storing energy in the bow and setting the restitution when the robot impacts with the ground. The orientation of the robot's body is passively stabilized by the placement of the leg pivot just above the body's center of mass. Zeglin and Brown have studied the problem of finding a sequence of hops that land on stepping stones [19]. The robot's two controls at each impact are the leg angle and the cocking of the leg (impact restitution). These control variables are analogous to the two controls in a bat by a one joint robot: the angle and angular velocity of the robot at impact.

Batting is a form of underactuated manipulation (Lynch and Mason [6,7]). In underactuated manipulation, a low-degree-of-freedom robot controls more part freedoms through dynamic coupling. The part necessarily moves relative to the manipulator (e.g., slipping, rolling, and other forms of *nonprehensile* manipulation). When the robot maintains continuous contact with the part, the infinite-dimensional robot trajectory can be varied to control the part's freedoms. In the case of planar batting, where we have discrete impact events, the part has three configuration variables and six state variables, and only two controls are available at each impact. Therefore, in general, at least three bats are required to transfer the part to a specific state.

1.1 Batting a Disk in the Plane

We examine manipulation by planar batting by a one joint robot in a gravitational field. The robot is dubbed 1JAG, for one joint and gravity.[1] Our goal is to show that a simple batting robot can control all six state variables of an object in the plane by a sequence of bats. We use a predictive impact model to find the state of the object after a bat or sequence of bats, then compare the "distance" of the final state to some predefined goal state. We use nonlinear optimization on the control space to minimize this final distance. This approach results in nearly globally convergent bat juggling to a stable batting cycle when the controls correspond to a single bat (the one-bat). The approach has also been extended to a three-bat scheme that can bat an object from nearly any initial state to nearly any final goal state. While these are simulation results, the control computations are performed in real-time, and we are currently implementing the controller on our experimental setup.

1.2 Definitions

We define the configuration space of the object in the plane as $C = SE(2) = R^2 \times S^1$. The configuration in the world frame F_w fixed to the robot origin is given by $\mathbf{q} = (x, y, \phi) \in C$. The state space of the object is the tangent bundle $TC = SE(2) \times R^3$. Object states in the world frame are denoted $\mathbf{z} = (\mathbf{q}, \dot{\mathbf{q}}) = (x, y, \phi, \dot{x}, \dot{y}, \dot{\phi}) \in TC$.

---

[1] We have affectionately dubbed it "Sammy" in honor of the local Chicago Cubs batting legend "Slammin'" Sammy Sosa.

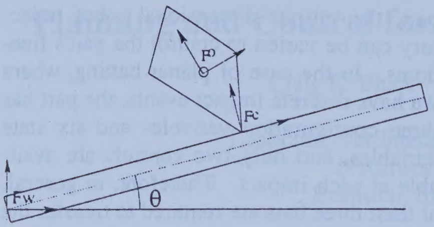


Figure 2
Definition and positions of coordinate frames.

A contact frame defined as F_c is attached to the batting arm at the point of contact. Object states in this frame are $\mathbf{z}_c = (x_c, y_c, \phi_c, \dot{x}_c, \dot{y}_c, \dot{\phi}_c)$. We also define an object frame F_o fixed to the object at the center of mass. Coordinates in this frame are given as (x_o, y_o, ϕ_o). See Figure 2.

The robot's state is $(\theta, \dot{\theta}) \in S^1 \times R^1$. These are measured in the world frame. All linear distances are given in centimeters, angles in radians, and times in seconds.

The batting function $\mathbf{z}_f = f(\mathbf{z}_0, \mathbf{u})$ calculates the object's final state $\mathbf{z}_f$ as a function of the initial state $\mathbf{z}_0$ and the control vector $\mathbf{u}$. For the simple one-bat, we take an object from an initial state $\mathbf{z}_0$, hit it once after time t_1 with batter angular velocity $\dot{\theta}_1$, and arrive at a final state $\mathbf{z}_1$ after t_2. This gives the one-bat control and its function.

$$\mathbf{u}_1 = (t_1, \dot{\theta}_1, t_2)$$

$$\mathbf{z}_1 = f_1(\mathbf{z}_0, \mathbf{u}_1) = f_1(\mathbf{z}_0, t_1, \dot{\theta}_1, t_2)$$

We can expand this to encompass a two-bat and a three-bat. The two-bat involves hitting an object two consecutive times. The three-bat is a sequence of three consecutive bats with seven control parameters (Figure 3).

$$\mathbf{u}_3 = (t_1, \dot{\theta}_1, t_2, \dot{\theta}_2, t_3, \dot{\theta}_3, t_4)$$

$$\mathbf{z}_3 = f_3(\mathbf{z}_0, \mathbf{u}_3) = f_3(\mathbf{z}_0, t_1, \dot{\theta}_1, t_2, \dot{\theta}_2, t_3, \dot{\theta}_3, t_4)$$

$$= f_1\left(\begin{matrix} f_1(f_1(\mathbf{z}_0, t_1, \dot{\theta}_1, \alpha t_2), (1-\alpha)t_2, \dot{\theta}_2, \beta t_3), \\ (1-\beta)t_3, \dot{\theta}_3, t_4 \end{matrix} \right)$$

where α and β are arbitrary values between 0 and 1.

We would like to find optimal control parameters $\mathbf{u}^*$ such that our final state is at

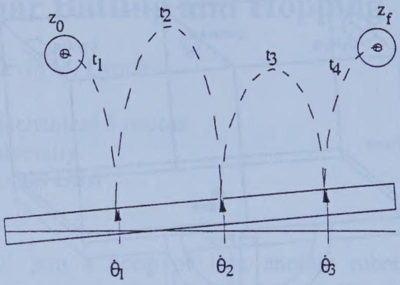


Figure 3
Parameters for sequentially batting an object in the plane.

the goal state, $\mathbf{z}_g = (x_g, y_g, \phi_g, \dot{x}_g, \dot{y}_g, \dot{\phi}_g) = f(\mathbf{z}_0, \mathbf{u}^*)$.

2. Batting

We model batting as the collision of rigid bodies with finite, nonzero friction and with a coefficient of restitution greater than zero and less than one. Friction is included in the model because it is present in the physical system, and it allows the control of orientation of a disk.

2.1 Impact of a Rigid Body

We have adopted Poisson's model of restitution as our initial model (Wang and Mason [18]). This model divides the collision into two stages, a compression state and a restitution state, each with accompanying impulses (P_{yc}, P_{yr}) normal to the impact surface. Poisson's restitution coefficient e is the ratio of these impulses. This method handles friction and slip while ensuring conservation of energy.[2] Poisson's hypothesis is identical to Newton's impact law if the collision is collinear[3] or if there is no friction.

---

[2] Stronge's [15] definition of restitution is the only one to give energetically consistent results if $e = 1$.

[3] A collinear impact refers to when the impact point and center of mass are collinear with the normal impact vector.

2.2 Batting Equation

For each bat, if we know the angle of the batter θ, the angular velocity of the batter $\dot{\theta}$, and the state of the object **z**, then we can uniquely determine the post impact state of the object. This is done by transforming the state of the object from the world frame F_w in which it is measured to the contact frame F_c. In F_c the impact is viewed as an interaction between a massive fixed surface (the batter) and a single moving object.

2.2.1 Free Flight, part 1

If $\mathbf{z}_0 = (x_0, y_0, \phi_0, \dot{x}_0, \dot{y}_0, \dot{\phi}_0)$ is our initial state at time 0, then the state at impact after time t_1 is given by the ballistic flight equations, where g is the gravitational constant acting in the -y direction in F_w.

$$x_1 = x_0 + \dot{x}_0 t_1 \qquad \dot{x}_1^- = \dot{x}_0$$

$$y_1 = y_0 + \dot{y}_0 t_1 + \tfrac{1}{2} g t_1^2 \qquad \dot{y}_1^- = \dot{y}_0 + g t_1$$

$$\phi_1 = \phi_0 + \dot{\phi}_0 t_1 \qquad \dot{\phi}_1^- = \dot{\phi}_0$$

2.2.2 Impact

The pre-impact state of the object is given by $\mathbf{z}_1^- = (x_1, y_1, \phi_1, \dot{x}_1^-, \dot{y}_1^-, \dot{\phi}_1^-)$ in the world frame F_w. In the contact frame F_c, the object velocity is

$$\dot{x}_c^- = (\dot{x}_1^- + \dot{\theta} r_{yc})\cos\theta + (\dot{y}_1^- - \dot{\theta} r_{xc})\sin\theta$$

$$\dot{y}_c^- = -(\dot{x}_1^- + \dot{\theta} r_{yc})\sin\theta + (\dot{y}_1^- - \dot{\theta} r_{xc})\cos\theta$$

$$\dot{\phi}_c^- = \dot{\phi}_1^- - \dot{\theta}.$$

Velocity in F_c can then be fed into the impact equations of motion from [18] to get the resulting impulses. These determine the post impact velocities in the contact frame. Assuming an object of unit mass, we have

$$\dot{x}_c^+ = \dot{x}_c^- + P_x$$

$$\dot{y}_c^+ = \dot{y}_c^- + P_y$$

$$\dot{\phi}_c^+ = \dot{\phi}_c^- + \left(\frac{P_x r_{yc} - P_y r_{xc}}{\rho^2} \right),$$

where ρ is the object's radius of gyration of inertia, P_x is the impulse in the x-direction, P_y is the impulse in the y-direction, and r_{xc} and r_{yc} are the x and y components of the vector from the contact point to the object center of mass.

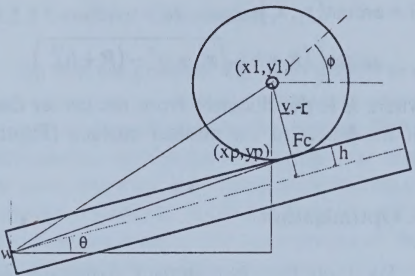


Figure 4
Impact geometry for a disk in the plane.

Transforming back to the world frame F_w, we get the post-impact velocities

$$\dot{x}_1^+ = \dot{x}_c^+ \cos\theta - \dot{y}_c^+ \sin\theta - \dot{\theta} r_{yc}$$

$$\dot{y}_1^+ = \dot{x}_c^+ \sin\theta + \dot{y}_c^+ \cos\theta + \dot{\theta} r_{xc}$$

$$\dot{\phi}_1^+ = \dot{\phi}_c^+ + \dot{\theta}.$$

This yields the post-impact state, given by $\mathbf{z}_1^+ = (x_1, y_1, \phi_1, \dot{x}_1^+, \dot{y}_1^+, \dot{\phi}_1^+)$.

2.2.3 Free Flight, part 2

The third control, t_2, picks a state along the object's post impact trajectory. The object is at the state $\mathbf{z}_2$ at a time t_2 after impact.

$$x_2 = x_1 + \dot{x}_1^+ t_2 \qquad \dot{x}_2 = \dot{x}_1^+$$

$$y_2 = y_1 + \dot{y}_1^+ t_2 + \tfrac{1}{2} g t_2^2 \qquad \dot{y}_2 = \dot{y}_1^+ + g t_2$$

$$\phi_2 = \phi_1 + \dot{\phi}_1^+ t_2 \qquad \dot{\phi}_2 = \dot{\phi}_1^+$$

2.3 Impact Geometry for a Disk

In the case that our object is a disk[4] of radius R, the vector $\mathbf{r}_c$ in F_c from the contact point to the object center of mass is always of length R and in the direction normal to the contact line, yielding

$$r_{xc} = 0$$

$$r_{yc} = R.$$

The angle θ of the batter impacting a disk at a configuration (x_1, y_1, ϕ_1) in F_w is

---

[4] We will consider batting polygonal objects in future work.

$$\theta = \arctan(y_1, x_1) - \arctan\left((R+h), \sqrt{x_1^2 + y_1^2 - (R+h)^2} \right)$$

where h is the distance from the center line of the batter to the contact surface (Figure 4).

3. Optimization

By modeling our impact dynamics we determine a "forward" batting function that predicts the final state of the object given its initial state and a set of batting controls. We can then use a nonlinear optimization scheme to find controls that minimize the final distance to the goal. Related approaches using nonlinear optimization for motion planning for underactuated systems appear in (Divelbiss and Wen [3]; Lynch and Mason [7]; Sussmann [16]). A solution to the minimization problem $\mathbf{u}^*$ represents a control set that takes the object near the goal state.

3.1 Objective Function

We define a metric measuring the distance of the final state to the goal state via a weighted objective function as follows:

$$\text{distance} = g(\mathbf{z}_f, \mathbf{z}_g) = g(f(\mathbf{z}_0, \mathbf{u}), \mathbf{z}_g)$$

$$g(\mathbf{z}_f, \mathbf{z}_g) = c_1(x_g - x_f)^2 + c_2(y_g - y_f)^2 + c_3\rho^2(\phi_g - \phi_f)^2 + c_4(\dot{x}_g - \dot{x}_f)^2 + c_5(\dot{y}_g - \dot{y}_f)^2 + c_6\rho^2(\dot{\phi}_g - \dot{\phi}_f)^2$$

where $c_1 \ldots c_6 \geq 0$ are objective weights that we may choose to our liking. Note that ϕ is mod(2π).

Figures 5 and 6 show contour plots for the objective function $g(\mathbf{z}_f, \mathbf{z}_g)$ in the (t_1, t_2) plane for a one-bat of a uniform mass disk. We fix the 1JAG impact velocity $\dot{\theta}_1 = 0.5$ rad/s.

gravity = -85.5 cm/s^2
$R = 3.0$ cm
$h = 3.0$ cm
$e = 0.7$
$\mathbf{z}_0 = (40.0, 40.0, 0.0, 5.0, 20.0, 10.0)$
$\mathbf{z}_g = (35.0, 35.0, 0.0, .0, 0.0, 2.0)$

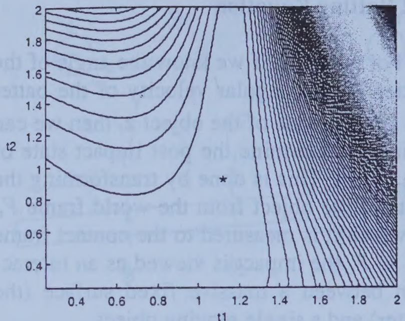


Figure 5
c = (4.0,4.0,0.0,0.5,0.5,0.0)
Objective function contour plot with zero weight on the angular variables. The minimum point is denoted (★).

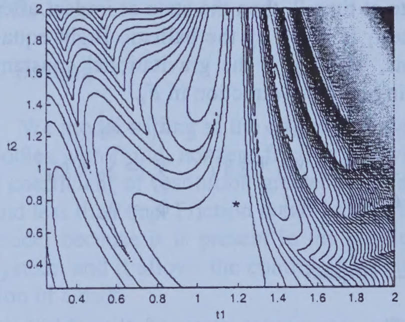


Figure 6
c = (4.0,4.0,400.0,0.5,0.5,0.5)
This objective function differs from that of Figure 5 by the weights on the angular variables. The ridges at the right of the plot demonstrate the effect of mod(2π) in the orientation. The minimum point is denoted (★).

3.2 Optimization Method

We use an unconstrained method for nonlinear optimization. Though our problem does have state and control variable constraints (negative flight times are impossible, batter velocities must be within range of the physical equipment, and large batting angles should be avoided), unconstrained methods are much faster and more likely candidates for real-time implementation in a control law. Solutions to the unconstrained optimization are only implemented if they satisfy inequality constraints such as those mentioned above.

These optimization schemes find *local* minimizers, not global minimizers. If the problem has many local minima, the solution will greatly depend on the initial starting point of the optimization. For this reason we consider a grid of initial guesses. A minimizer with a zero objective value exactly solves the batting problem.

3.2.1 BFGS Quasi-Newton Method

We use a quasi-Newton nonlinear optimization that calculates a step direction based on the gradient and an approximation to the Hessian. It then performs a line search along that direction to minimize the objective function, generating a new iterate. The optimization stops if the objective value of the new iterate is sufficiently close to the previous value. The Hessian approximation is generated using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. This method has the advantage of having super-linear convergence rate without needing to fully calculate and invert the Hessian.

BFGS Quasi-Newton Method [9]:

- Choose a search direction $\mathbf{P}_k = -[\mathbf{B}_k]^{-1}\nabla g(\mathbf{u}_k) = -\mathbf{H}_k \nabla g(\mathbf{u}_k)$ at point $\mathbf{u}_k$.
- Find a step length α_k via a line search that determines how far to move along $\mathbf{P}_k$.
- Calculate next point $\mathbf{u}_{k+1} = \mathbf{u}_k + \alpha_k \mathbf{P}_k$.
- Stop when $\|g(\mathbf{u}_{k+1}) - g(\mathbf{u}_k)\|$ is small.

BFGS updating formula:

$$[\mathbf{B}_k]^{-1} = \mathbf{H}_k \approx [\nabla^2 g(\mathbf{u}_k)]^{-1}$$

$$\mathbf{s}_k = \mathbf{u}_{k+1} - \mathbf{u}_k$$

$$\mathbf{y}_k = \nabla g(\mathbf{u}_{k+1}) - \nabla g(\mathbf{u}_k)$$

$$\rho_k = \frac{1}{\mathbf{y}_k^T \mathbf{s}_k}$$

$$\mathbf{H}_{k+1} = (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^T)\mathbf{H}_k(\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^T) + \rho_k \mathbf{s}_k \mathbf{s}_k^T$$

3.2.2 Gradient Calculation

To find the gradient ∇g with respect to a control vector **u**, we have chosen to use a numerical approximation by central finite differences. Analytical derivatives can also be used, but they are highly complex trigonometric functions which are slow to evaluate. Another issue is the smoothness of the batting function at the boundaries between different impact types (sliding, C-sticking, etc. [18]).

4. Planner

The planner finds controls that minimize the objective function $g(\mathbf{z}_f, \mathbf{z}_g)$, filters out infeasible controls, and determines how to process the chosen solution.

4.1 Optimization Seeds

The objective function g is complex and tends to contain multiple minima. By choosing a dense grid of initial guesses in the control space and optimizing from each point, we can be fairly certain of finding the globally optimal control. This is in conflict, however, with our eventual requirement that the computations be performed in real-time. As a compromise, we perform a single objective function evaluation for a grid of initial points, and choose the point with the lowest objective value as our initial guess. This allows us to perform a rough search of the solution space without spending the time to optimize from each point in the grid.

We place constraints on the control space to define a feasible region. Times must be greater than some small positive threshold value. Angular velocities must be within motor limits, though typically we enforce much tighter constraints.

4.2 Bat Sequences

4.2.1 The One-bat

The one-bat optimization involves batting an object once in order to bring it as close as possible to the goal state. For the one-bat we have only three controls, given by $\mathbf{u}_1 = (t_1, \dot{\theta}, t_2)$. In general this is not sufficient to get exactly to a goal state. To specify an arbitrary point in the six-dimensional state space we need at least six control variables.

For this reason, we use the one-bat to reach given (x,y) positions and velocities. The one-bat is sufficient for stable periodic juggling to a fixed point from a large set of initial states. Angular weights c_3 and c_6 are 0 in the objective function $g(\mathbf{z}_f, \mathbf{z}_g)$.

4.2.2 The Three-bat

The three-bat has seven controls, allowing the robot to control all six state variables of the object (within inequality constraints) with one redundant control. By optimizing parameters for three consecutive bats, we are not limited to be near any point at any intermediate time. This allows the planner to take the object far away from the goal in order to reorient or spin the object to match both orientation and rotational velocity. We find that the three-bat is capable of taking the object exactly to some goal state when the objective function is fully weighted for $(x, y, \phi, \dot{x}, \dot{y}, \dot{\phi})$.

4.3 Three-bat Planning Heuristic

One problem with the three-bat alone is that the process of finding a legitimate batting sequence is lengthy. This is a seven variable function as opposed to the three variable function for a one-bat. It takes about 25 seconds to search and optimize the three-bat, whereas the one-bat typically runs in 5-10 milliseconds. In the case of the three-bat, the vast majority of time is spent on function evaluations of the grid of initial control guesses, which typically consists of 10-20 points along each of four time dimensions and 5-10 points along each of three velocity dimensions.

To facilitate real-time three-bat planning, we keep the disk in a stable one-bat juggle while the planner searches for a three-bat sequence. As a heuristic, the planner generates a random stationary (x,y) subgoal in a neighborhood of the true goal and finds a one-bat plan. This generates solutions $\mathbf{u}_1^*$. We extrapolate the batting parameters for the one-bat into three-bat controls,

$$\mathbf{u}_1^* = (t_1, \dot{\theta}_1, t_2),$$

$$\mathbf{u}_3^0 = (t_1, \dot{\theta}_1, 2t_1, \dot{\theta}_1, 2t_2, \dot{\theta}_1, t_2)$$

and then use $\mathbf{u}_3{}^0$ as the seed for the three-bat optimization. This gives a fast technique for finding a "reasonable" three-bat seed while maintaining a stable cyclic trajectory for the disk.

4.4 Calculating the 1JAG Trajectory

We use a third-order polynomial trajectory to connect the current state $(\theta_0, \dot{\theta}_0)$ at time $t = 0$ to the desired impact state $(\theta_1, \dot{\theta}_1)$ at time t_1:

$$\theta(t) = a_3 t^3 + a_2 t^2 + a_1 t + a_0 .$$

This gives four variables for our four boundary conditions.

$$\theta(0) = \theta_0 \qquad \dot{\theta}(0) = \dot{\theta}_0$$

$$\theta(t_1) = \theta_1 \qquad \dot{\theta}(t_1) = \dot{\theta}_1$$

Solving, we get the polynomial coefficients

$$a_0 = \theta_0 ,$$

$$a_1 = \dot{\theta}_0 ,$$

$$a_2 = \frac{-2\dot{\theta}_0 t_1 - \dot{\theta}_1 t_1 - 3\theta_0 + 3\theta_1}{t_1^2} ,$$

$$a_3 = \frac{\dot{\theta}_0 t_1 + \dot{\theta}_1 t_1 + 2\theta_0 - 2\theta_1}{t_1^3} .$$

Plugging in at the current time we calculate θ_d, $\dot{\theta}_d$ and $\ddot{\theta}_d$, which are fed into a 1JAG PD controller with a feedforward acceleration term.

4.5 Feedback Control

Immediately after a bat, the controller calculates a new one-bat or three-bat control $\mathbf{u}^*$ based on the estimated state of the disk using the methods described above. To account for noisy data and perturbations in the flight of the disk, the controller updates $\mathbf{u}^*$ at each control tick as new state data becomes available. In the case of the one-bat, this is done by optimizing again, using the solution at the previous cycle as the seed to the optimization. Since a grid of initial guesses is not searched (as with the initial solution), optimization is quick and can occur in a single control cycle (1 ms in our Flatland system). Provided perturbations from the idealized flight dynamics are small, the change in the optimal control, and therefore the planned 1JAG trajectory, should be small. This assumes sufficient smoothness of the objective function.

In the case of a three-bat plan, an intermediate goal state in the three-bat plan is chosen as a subgoal for a one-bat optimization. This attempts to bring the system back to the originally planned object trajectory. If errors are too large (the objective function of the minimizer is large), the system must replan the three-bat.

5. Simulation Results

To test the planner, a simulator was written in C++ to simulate our experimental setup using the same impact analysis as in the controller. This section presents results of the simulation of the controller.

5.1 Setup

All simulations were run under a -85.5 cm/s^2 gravitational field, corresponding to a 5 degree tilt of our air table. We use a restitution coefficient of 0.7 and a friction coefficient of 0.25. The disk has radius R = 3.0 cm and the batter has an offset h = 3.0 cm. The goal state is $\mathbf{z}_g = (35.0,35.0,0.0,.0,0.0,2.0)$.

5.2 The One-bat

Our one-bat stable juggling succeeds in cyclic batting of a disk to a fixed (x,y) point. Recall this one-bat ignores rotational parameters. We find this is robust to perturbations and has a large domain of attraction. Results are shown in Figures 7 and 8.

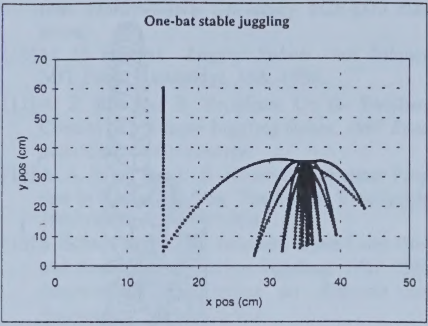


Figure 7
Convergence of the one-bat from a distant point $\mathbf{z}_0 = (15.0,60.0,0.0,0.0,0.0,0.0)$.

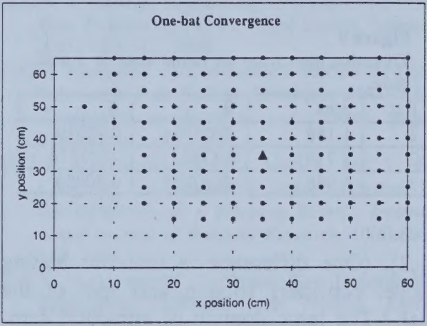


Figure 8
Domain of attraction for the one-bat. The disk is dropped from rest. All initial positions • converge to the stable fixed point ▲. Points were sampled within experimental bounds every 5 cm. Points lower than 6 cm have initial contact with the batter, and are not valid.
The empty region to the left (small x values) is a result of the chosen initial search grid for the optimization. The minimizers here are often infeasible. An alternate plan of choosing "second best" but feasible solutions sometimes results in convergent controls.

We use objective weights given by $\mathbf{c} = (4.0,4.0,0.0,0.5,0.5,0.0)$. Our initial search range over which we sample function values is $t_1 = 0.5:0.05:1.5$, $\dot{\theta}_1 = 0.0:0.1:1.2$, $t_2 = 0.5:0.05:1.5$. We define the feasible region as $t_1 > 0.2$, $-0.5 < \dot{\theta}_1 < 3.0$, $t_2 > 0.2$.

With the one-bat we ignore the angular variables, and therefore the goal is the same as that of Bühler and Koditschek's juggler

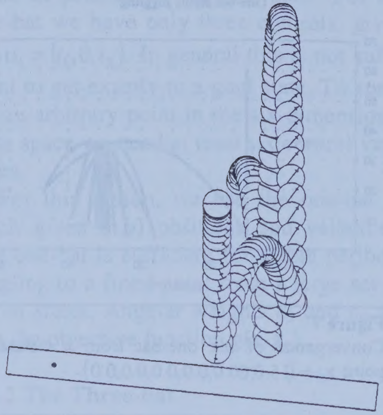


Figure 9
Three-bat plan from an initial state to a goal state.

Bat	time	θ	$\dot{\theta}$
1	1.189	-0.05764	0.70019
2	3.710	0.11841	-0.27374
3	4.988	-0.10128	0.50980

[1,2]. (One difference is that our batting model considers friction and spin of the disk.) The large domain of attraction demonstrated in simulation is reminiscent of the mirror law. In future work we plan to study the relationship between these two controllers. We note that the physical extent of the disk and the robot are explicitly modeled in the impact equations in our planner, but not in the mirror law.

5.3 The Three-bat

Using the three-bat planning technique, we can plan a sequence of bats that takes the disk exactly[5] to the goal state including angular variables.

The three-bat example in Figures 9 - 11 took about 25 seconds to find an optimization seed from a 7-dimensional grid and 75 iterations from that point. Optimization time is a few milliseconds.

$\mathbf{c} = (4.0, 4.0, 400.0, 0.5, 0.5, 0.5)$

$\mathbf{z}_0 = (40.0, 40.0, 0.0, 5.0, 20.0, 10.0)$.

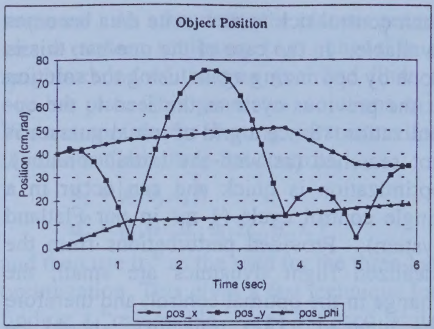


Figure 10
Object position during the three-bat. Note convergence to the goal state. Final φ position is zero (mod(2π)).

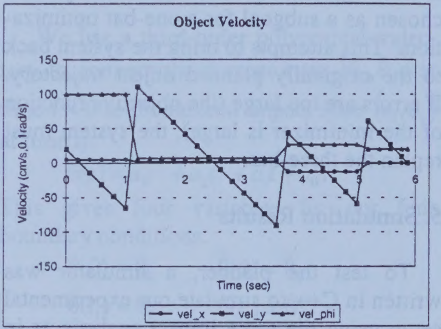


Figure 11
Object velocity during the three-bat.

$\mathbf{u}_3 = (1.19, 0.70, 2.52, -0.27, 1.28, 0.51, 0.87)$

final objective value = 3.447e-14

6. Discussion

This paper describes preliminary work on deriving a planner and controller that can be applied to both manipulation by batting and locomotion by hopping. The solution method we have adopted is a type of model predictive control and is quite general; the same idea can be applied to any system for which we have a model of dynamics. This generality is also the weakness of the approach – we are not taking advantage of any system-specific structure to ensure convergence or stability of the system. Ideally our controller would be guaranteed to (1) find a batting plan when one exists and (2) perform its computations in a fixed amount of time. We are encouraged by existing analytical results for simplified batting and hopping [1,2,5,8,17] and hope to find similar results for a simplified version of the dynamics (point mass, zero friction, or line batter) and simplified optimization method.

---

[5] Objective function returning 10^{-9} or less.

Another challenge is implementing the controller on our experimental testbed Flatland. Initial tests are underway. It remains to be seen how well the impact model predicts the behavior of the system, and how well discrepancies between the actual system and the model can be compensated by the feedback control described in Section 4.5.

In the future, we plan to apply the controller to batting manipulation of polygons and hopping robots similar to Brown and Zeglin's [19].

References

[1] M. Bühler and D. E. Koditschek. From Stable to Chaotic Juggling: Theory, Simulation, and Experiments. *1990 IEEE International Conference on Robotics and Automation*, pages 1976-1981.

[2] M. Bühler, D. E. Koditschek and P. J. Kindlmann. Planning and Control of Robotic Juggling Tasks. *Fifth International Symposium on Robotics Research*, pages 270-281, Tokyo, Japan, 1989.

[3] A. W. Divelbiss and J. Wen. A Global Approach to Nonholonomic Motion Planning. *Proceedings of the 31st Conference on Decision and Control*, pages 1597-1602, December 1992.

[4] J. Hodgins and M. H. Raibert. Biped Gymnastics. *International Journal of Robotics Research*, Vol.8, 1990.

[5] D. E. Koditschek and M. Bühler. Analysis of a Simplified Hopping Robot. *International Journal of Robotics Research*, 10(6):587-605, December 1991.

[6] K. M. Lynch and M. T. Mason. Dynamic Manipulation with a One Joint Robot. *1997 IEEE International Conference on Robotics and Automation*, pages 359-366.

[7] K. M. Lynch and M. T. Mason. Dynamic Nonprehensile Manipulation: Controllability, Planning, and Experiments. To appear, *International Journal of Robotics Research*.

[8] R. T. M'Closkey and J. W. Burdick. Periodic Motion of a Hopping Robot with Vertical and Forward Motion. *International Journal of Robotics Research*, 12(3):197-218, 1993.

[9] J. Nocedal and S. Wright. Numerical Optimization. Northwestern University ECE-D79 class notes.

[10] M. H. Raibert. *Legged Robots That Balance.* MIT Press, Cambridge, MA, 1986.

[11] A. Z. Río and B. Brogliato. On the Feedback Control of a Simple Juggling Robot. *1997 European Control Conference.*

[12] A. A. Rizzi and D. E. Koditschek. Further Progress in Robot Juggling: The Spatial Two-Juggle. *1993 IEEE*, pages 919-924.

[13] S. Schaal and C. G. Atkeson. Open Loop Control Strategies for Robot Juggling. *1993 IEEE International Conference on Robotics and Automation*, pages 913-918.

[14] M. W. Spong. Development of a Three Degree of Freedom Air Hockey Robot. Video proceedings, *1998 IEEE International Conference on Robotics and Automation.*

[15] W. J. Stronge. Rigid Body Collisions with Friction. *Proceedings of the Royal Society*, London, A431:169-181, 1990.

[16] H. Sussmann. A Continuation Method for Nonholonomic Path-Finding Problems. *1993 IEEE Conference on Decision and Control*, pages 2718-2723.

[17] A. F. Vakakis, J. W. Burdick, and T. K. Caughey. An Interesting Strange Attractor in the Dynamics of a Hopping Robot. *International Journal of Robotics Research*, 10(6):606-618, 1991.

[18] Y. Wang and M. T. Mason. Two Dimensional Rigid-Body Collisions with Friction. *ASME Journal of Applied Mechanics*, 59:635-641, Sept. 1992.

[19] G. Zeglin and B. Brown. Control of a Bow Leg Hopping Robot. *1998 IEEE International Conference on Robotics and Automation*, pages 793-798.

[20] N. B. Zumel and M. A. Erdmann. Balancing of a Planar Bouncing Object. *1994 IEEE International Conference on Robotics and Automation*, v. 4, pages 2949-2954.

Stabilization of Balance in 3-D Bipedal Locomotion

A. D. Kuo, C. E. Bauby
Dept. of Mechanical Engineering & Applied Mechanics, University of Michigan, Ann Arbor, MI 48109-2125

ABSTRACT

We developed a model of passive dynamic walking incorporating lateral motion, and found that the forward modes are stable while one lateral mode is unstable. We considered several methods for stabilizing this mode, including direct control of angular momentum in the roll plane, as well as an indirect method involving lateral adjustment of stepping position. We found that lateral foot placement can stabilize the system with arbitrarily low cost. Stability is achieved by adjusting the amount of dissipation occurring at each step, transferring a controlled amount of energy to or from the lateral mode. This method is in contrast to direct methods, which must provide or dissipate energy in amounts comparable to those associated with a perturbation. We then hypothesized that the human central nervous system (CNS) must actively stabilize lateral motion but not necessarily forward motion, and that lateral stability is achieved using lateral foot placement. As a simple test, we measured variability of foot placement during natural gait for subjects walking with eyes open and eyes closed, to see if lateral variability is increased more than fore-aft variability when eyes are closed, as would be predicted by the hypothesis.

REVIEW AND THEORY

McGeer (1990) showed that a simple two-legged mechanism can walk stably down a gentle slope with no external control or energy input. His model was a coupled pair of pendula restricted to the plane. Given proper initial conditions, the system exhibited a walking limit cycle with asymptotic local stability to perturbations. We extended the planar passive dynamic walker to three dimensions, and considered the case in which the machine can rock side to side.

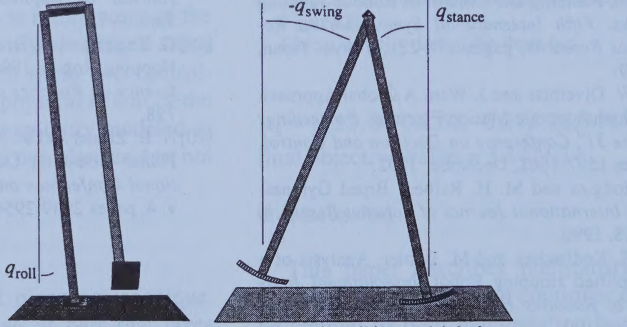


Figure 1. Three-dimensional passive dynamic walking model. Left: Front view. Right: View from right.

Of particular interest is the control of human walking. Balance is maintained during human walking through the interaction of CNS control and body dynamics. Several investigators (cf. McKinnon & Winter, 1993, Redfern & Schumann, 1994) have hypothesized that control of lateral balance is of particular importance, based on evidence such as the high incidence of sideways falls in the elderly (Hayes et al., 1996). There is, however, little understanding of which components of gait are inherently stable, and which require active control.

The three-dimensional passive dynamic walking model consists of two legs with curved feet which are geometrically modeled as sections of cylinders (see Fig. 1). There are three degrees of freedom. The stance and swing legs rotate about an axis through the hips, with angles q_{stance} and q_{swing} measured counter-clockwise from the upright position. The stance leg also rotates laterally in the frontal plane about another pin joint between the stance foot and leg, whose axis runs in the fore-aft direction when the machine is in the upright position. This roll angle, q_{roll}, is measured counter-clockwise from the upright position viewed from the front. This system will act as two coupled pendula, with the swing leg swinging forward until it makes contact with the ground. The equations of motion are written in standard robotics form as

$$M(q)\ddot{q} = g(q) + v(q,\dot{q}) \qquad (1)$$

where $q \equiv [q_{\text{roll}}\ \ q_{\text{stance}}\ \ q_{\text{swing}}]^T$, M is the mass matrix, g is a vector containing gravitational terms, and v is a vector containing Coriolis and centripetal terms. These equations are nonlinear and were solved numerically; a Runge-Kutta algorithm was used to calculate the solution to the corresponding initial-value problem,

$$\dot{x} = f(x_0) \qquad (2)$$

where the state vector $x \equiv \begin{bmatrix} q^T & \dot{q}^T \end{bmatrix}$ and x_0 is an initial condition in which the machine is in the double-support position.

A full swing comprises the motion of the machine starting from the initial double-support position, including approximately one half-period swing of the swing leg, and ending when the swing foot contacts the ground. This condition occurs when the distance from the ground of the bottom of the swing foot reaches zero:

$$g(x) = 0 \qquad (3)$$

A small angle approximation is employed for q_{roll} at ground contact, because for mechanisms with non-zero pelvis width w, both feet can make perfect line contact with the ground only when $q_{\text{roll}} = 0$ and $q_{\text{stance}} = -q_{\text{swing}}$. It is assumed that contact only occurs when the machine is near that configuration with small q_{roll}. The state of the system following one swing is denoted by the '-' superscript, e.g., x^-.

Assuming a perfectly inelastic collision and an instantaneous transfer of weight from the trailing leg to the leading leg, the states after impact are found through conservation of angular momentum. Three conservation equations are required to determine the velocities $\dot{q}$ following ground contact. One component of angular momentum of the entire machine is conserved about the ground contact point, along an axis transverse to the foot. Another component is conserved about the hinge joint on the foot, along the axis of the joint. Finally, one component is conserved for angular momentum of the trailing leg about the pelvis, along an axis running through the hips. These equations may be summarized into a single matrix equation

$$\dot{q}^+ = L(q^-)\dot{q}^- \qquad (4)$$

After impact, the trailing leg becomes the new swing leg, and the leading leg becomes the new stance leg. A necessary condition for steady walking is that the system be capable of a repetitive motion (limit cycle). Writing the initial conditions for a step $k+1$ as a function of the previous step $F(x_k)$, we require

$$x_{k+1} = F(x_k) = x_0 \tag{5}$$

The search for the appropriate conditions (a fixed point) involves the solution of a two-point boundary value problem. We employed a simple first-order shooting method search for the fixed point located at double-support. There are two solutions, one which has two unstable modes and has a short time of swing, and another which has one unstable mode and has a longer time of swing. We will be concerned with the long-period gait, because the instability is isolated to the lateral motion.

Stability and Control

We found that the passive walking in three dimensions is unstable, as indicated by the

eigenvalues	-33.335	-0.958	0.375	-0.025	-0.003
roll angle	0.682	-0.049	-0.008	-0.709	-0.323
stance angle	0.238	-0.866	-0.736	-0.009	0.352
roll velocity	0.673	0.015	0.012	0.702	0.333
stance velocity	0.160	0.403	0.607	0.023	-0.216
swing velocity	-0.001	-0.291	-0.300	0.066	-0.784

Table 1. Eigenvalues & eigenvectors of step-to-step transition matrix (Poincaré map). Components of associated mode shown below each eigenvalue.

magnitude of an eigenvalue of the discrete long-period step-to-step map exceeding unity (see Table 1). The results show that the unstable mode is primarily confined to the lateral states, meaning that the forward motion remains passively stable as in the planar model. Feedback –control is necessary to stabilize this mode.

There are several means by which actuation can be added to the passive machine to achieve stabilization. We first consider direct methods, which seek to control angular momentum in the roll plane so that the state trajectory is steered towards the fixed point following a disturbance. The most direct method to actuate angular momentum would be to incorporate a reaction wheel at the pelvis. In mid-stride, commanding motion of the reaction wheel causes angular momentum of the passive machine (less the reaction wheel) to change, because total angular momentum must be conserved about the end of the stance leg. Another direct method is to install an actuated inverted pendulum on the pelvis. In this method, steering of the trajectory is achieved by shifting mass to one side or another. Rather than carefully analyze either of these two methods, we recognize that there is a simple lower bound on the cost needed for stability. At the start of a step, a perturbation affects roll velocity, because roll angle is constrained by double support conditions. This velocity error must be countered by a motion of the reaction wheel or inverted pendulum such that the state is restored to the nominal trajectory. A (conservative) lower bound on the cost of direct method stabilization, expressed in terms of an impulse or change in position of the center of mass, is equal to the impulse or change in position caused by the perturbation.

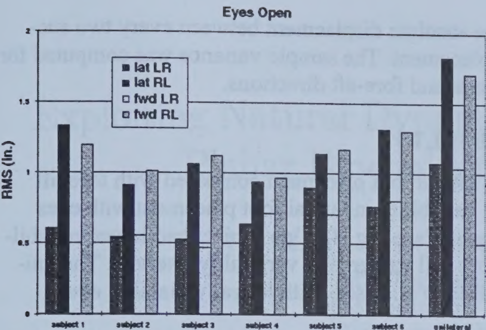


Figure 3: Variance of foot placement for seven normal subjects (and one with unilateral vestibular deficit), eyes open trials. Shown are forward and lateral variability, for left-to-right foot and right-to-left footsteps. Lateral variability was larger in all subjects.

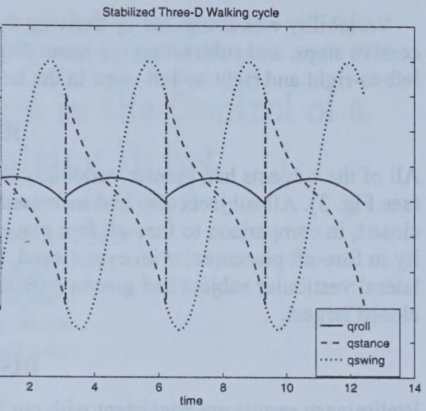


Fig. 2. Roll, stance, and swing angles for four steps of a stabilized gait cycle

Another way to control the roll mode is by the indirect method of lateral foot placement. By extending the lateral distance between the feet quasi-statically during a stride, it is possible to adjust the amount of energy dissipated or transferred from the forward modes at ground contact. A simple analysis of the linearized discrete step-to-step mapping verifies that lateral foot placement yields a controllable system. Placing the unstable eigenvalue at 0 results in gains dominated by proportional-derivative gains of 1 for q_{roll} and 1 for $\dot{q}_{\text{roll}}$. In fact, setting the other gains to zero still results in a stable system. This is due to the amount of decoupling between roll and forward dynamics. The trajectory of the configuration variables for the stabilized system are shown in Fig. 2.

EXPERIMENT

Based on the theoretical results for the model, we hypothesized that humans may be subject to the same dynamics, implying that lateral stability is a major concern, but not necessarily fore-aft stability. A simple test of our hypothesis is to compare relative placement of the feet while walking with eyes open and with eyes closed. If the CNS actively stabilizes lateral balance, loss of vision should result in greater variability in lateral foot placement. Because the forward walking modes appear to be passively stable, the CNS need not actively control them, and loss of vision should result in little or no changes in variability of forward foot placement.

We measured lateral foot placement in the natural gait of seven normal adult subjects and one with unilateral vestibular damage, aged 18-40. Foot and other body segment kinematics were measured using a Flock of Birds (Ascension Technology Inc., VT) magnetic tracking system mounted on a rolling cart. The cart was pushed behind subjects who walked at a naturally selected speed in a subjectively determined straight line. The absolute location of the cart was measured using two optical encoder wheels rolling on the ground, making it possible to reconstruct absolute position of the feet. When subjects walked with eyes closed, their trajectory was guided by having them follow a guide providing a voice cue by speaking continuously. All subjects walked at least 100 contiguous steps in a trial.

Variability was computed by deriving the absolute displacment between every two successive steps, and subtracting the mean displacement. The sample variance was computed for left-to-right and right-to-left steps in the lateral and fore-aft directions.

RESULTS

All of the subjects had greater variability in lateral foot placement compared with fore-aft (see Fig. 3). All subjects also had increased variability in lateral foot placement with eyes closed, in comparison to fore-aft foot placement (see Fig. 4). One subject had lower variability in fore-aft placement with eyes closed, but still had greater variability laterally. The unilateral vestibular subject had greater variability in all cases, with lateral variability eyes closed largest.

DISCUSSION

Preliminary results are consistent with our hypothesis that lateral balance is actively controlled by the CNS, and also imply that vision plays a role in that control. There were also smaller increases in variability in fore-aft placement which may indicate a visual role in forward modes as well, but these differences may be due to the fact that subjects walked at a less steady speed with eyes closed. The difference between fore-aft and lateral variability does suggest that the inherently stable passive dynamics of the legs in the passive walking mechanism may be emulated by the biomechanics and local reflexes of the human lower extremity.

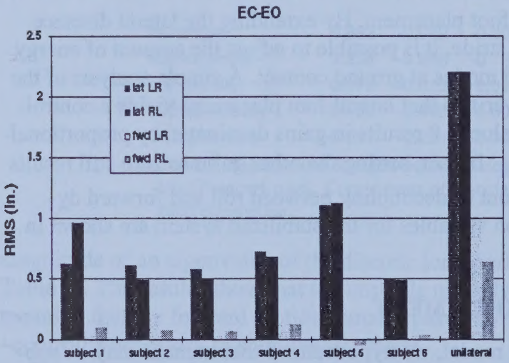


Figure 4: Variance of foot placement for same subjects, shown as difference between eyes closed and eyes open values. Values shown are lateral left-to-right and right-to-left variability, followed by fore-aft left-to-right and right-to-left variability. For each subject, compare first two columns with second two.

REFERENCES

MacKinnon CD & Winter DA (1993) J. Biomech 26: 633-44
Redfern MS, Schumann, T (1994) J. Biomech 27:1339-1346
Hayes WC et al. (1996) Bone 18: 77S-86S
McGeer T (1990) Int J Robot Res 9: 68-82

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant 1R29DC02312-01A1, the Whitaker Foundation, and NSF grant IBN-9511814.

Exploiting Natural Dynamics in the Control of a Planar Bipedal Walking Robot

Jerry E. Pratt and Gill A. Pratt
MIT Leg Laboratory
545 Technology Square
Cambridge, MA 02139
www.leglab.ai.mit.edu

Abstract

Natural dynamics can be exploited in the control of bipedal walking robots: the swing leg can swing freely once started; a kneecap can be used to prevent the leg from inverting; and a compliant ankle can be used to naturally transfer the center of pressure along the foot and help in toe off. Each of these mechanisms helps make control easier to achieve and results in motion that is smooth and natural looking.

We describe a simple control algorithm using these natural mechanisms which requires very little computation. The necessary sensing consists of joint angles and velocities, body pitch and angular velocity, and ground reaction forces.

Using this simple algorithm, we have controlled a seven link planar bipedal robot, called Spring Flamingo, to walk. Video, photographs, and more information on Spring Flamingo can be found at http://www.leglab.ai.mit.edu

1 Introduction

A powerful practice in machine design and control is to design mechanisms which have natural dynamics that make control simpler and more efficient. For example, planes have wings so that they naturally fly stably, requiring only a simple power source and simple control. Early locomotives used fly-ball governors, a mechanical feedback device, to help maintain constant speed. Satellites and rifle bullets spin to help maintain orientation.

Nature has evolved creatures that exploit the natural dynamics of similar mechanisms. Birds have wind-swept wings. Fish are neutrally buoyant. Running animals have springy legs. And walking animals have kneecaps, compliant ankles, and legs which can swing passively.

Many researchers [5, 10, 4, 1, 3], starting with Mc Geer and his passive dynamic walker, have exploited natural dynamics in order to make walking machines which are fully passive. These devices rely completely on their dynamics, and interaction with gravity, in order to walk.

Passive walkers have limitations, of course, such as limited capabilities and the need to walk down a slope. Powered robots [6, 16, 15, 14, 2, 8, 9, 11, 18, 19] can avoid these limitations. However, the control of powered bipedal robots has often been very complicated and the resultant motion often looks unnatural and is inefficient. Many of the controllers for powered robots are model based, requiring an accurate model of the dynamics of the

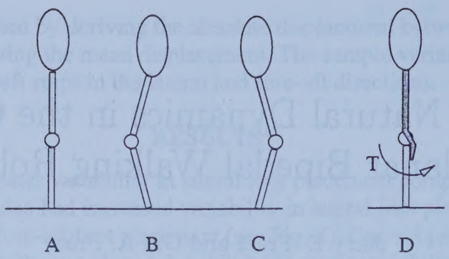


Figure 1: Diagram illustrating kneecap advantages. Without a kneecap, a biped with a straight support leg is in an unstable buckling configuration (A). Feedback control will result in chatter between knee inflections B and C due to delay, etc. With a kneecap (D), a constant torque with no feedback is enough to stablize the system against buckling.

robot in order to work. Several of the robots use trajectory tracking approaches which require pre-specified trajectories of either the body or the joints themselves.

In this paper, we describe an approach to powered bipedal walking which exploits the natural dynamics of the robot and requires only simple control. We exploit three different natural mechanisms. We use a knee cap to prevent the leg from inverting, which makes control of height easy. We use a compliant ankle limit so that the center of pressure on the foot travels forward with the center of mass of the body. And we exploit the natural swing dynamics of the leg to make swing control simpler and natural looking. This approach was used in both simulation and on a physical robot as described below. The robots walk fairly naturally and efficiently.

2 Natural Dynamic Mechanisms

2.1 Knee Cap

Walking with straight support legs is more efficient than with bent legs since energy requirements in muscles and motors are proportional to the torque at the joint, even if there is no velocity. However, since the leg must support the weight of the body, a straight leg poses an interesting challenge. Figure 1 illustrates the issue. When the body is directly over the foot (A), no torque is required at the knee. However, this is an unstable latch configuration. If the knee moves slightly either way, the leg will buckle (B or C). It is challenging to control this situation. Due to controller non-idealities (bandwidth limitation, delays, etc.) a straight knee controller will typically exhibit chatter between configurations B and C.

Adding a knee cap (D) can greatly simplify the control and make the resultant motion smoother and more efficient. A very simple control technique to keep the leg straight is to apply a constant torque so that the knee pushes against the stop. Of course other techniques can be used. Also, if the line of force on the body passes in front of the kneecap, the knee will be locked against the kneecap without any actuator torque.

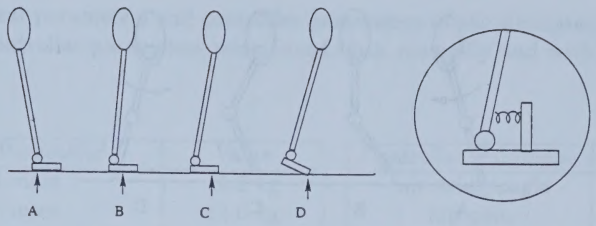


Figure 2: Diagram illustrating compliant ankle. In normal walking, the center of pressure on the foot travels forward as the center of mass travels forward (A-D). A compliant ankle (insert) can naturally achieve this effect. However, energy injection at toe off requires actuation.

2.2 Compliant Ankle

Feet and ankles provide many benefits to bipedal walking. They reduce velocity fluctuations since the center of pressure on the foot can travel forward, staying below the center of mass of the body. They also help to control speed and to inject energy at the end of the stride through toe off.

The torque at the ankle can be controlled actively. However, torque requirements can be quite high, since the foot provides a significant lever arm when the center of pressure is near the toe. A compliant ankle provides most of the benefits of a foot and ankle but without the torque requirements. An actuator can then be used in addition to the passive ankle for fine control and energy injection at toe off. Figure 2 illustrates the situation. In configuration A, the center of mass is behind the foot and there is zero ankle torque. In configurations B and C, the center of mass is travelling forward. The ankle torque increases, thereby moving the center of pressure of the foot forward from the heel to the toe. In configuration D, the robot goes into toe off, releasing the energy stored in configurations B and C and perhaps injecting some more, through active torques, to maintain walking. The inset shows a simple spring configuration which can give the ankle the desired compliance.

Choosing an adequate spring torque versus displacement curve is important in achieving the desired behavior. In the simulation discussed below we use a quadratic spring ($\tau = k(\theta - \theta_0)^2$) and tune the stiffness parameter. For the robot we use a rubber stop and adjust the position of the stop for best results.

2.3 Passive Swing Leg

In our previous algorithm for Spring Flamingo [15], as with most powered bipedal walkers, we used control techniques similar to those used for robotic arms to control the swing leg along a trajectory to a desired landing position. However, with a suitable leg, the natural swing dynamics are such that once the swing starts, the leg will continue without any intervention, as illustrated in Figure 3. Gravity alone can be used to initiate swing, as in the case of the passive dynamic walkers. Hip torque can be added in order to make the leg swing faster as in animals.

We use the passive swing properties of the leg in the control of our simulation and robot. In both cases, the hip is servoed forward to a desired angle and the knee is allowed to swing freely. At the end of the swing, moderate damping is added to the knee

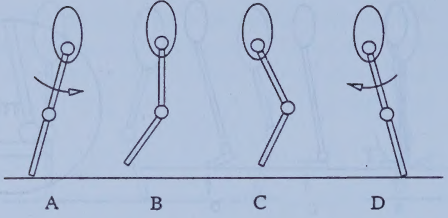


Figure 3: Diagram illustrating passive swing. Swing is initiated (A) through a forward torque on the hip, supplied either by hip actuators or gravity. The leg can swing passively (B - C) until swing is stopped (D) through a backward torque on the hip, again supplied either by hip actuators or gravity.

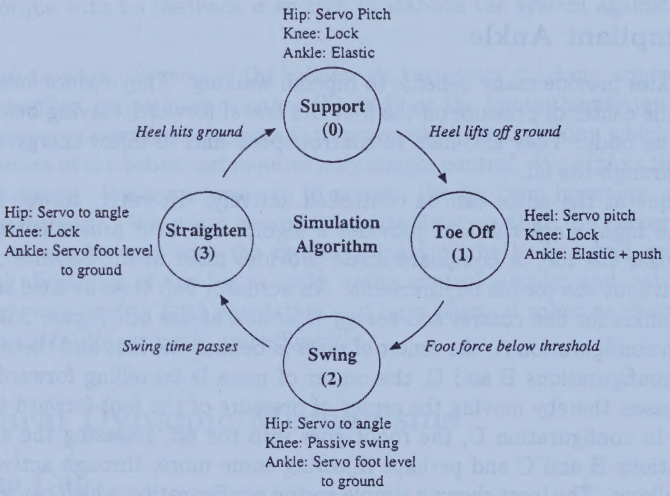

Hip: Servo Pitch
Knee: Lock
Ankle: Elastic
Support
(0)
Heel hits ground
Heel lifts off ground
Hip: Servo to angle
Knee: Lock
Ankle: Servo foot level to ground
Straighten
(3)
Simulation Algorithm
Toe Off
(1)
Heel: Servo pitch
Knee: Lock
Ankle: Elastic + push
Swing time passes
Foot force below threshold
Swing
(2)
Hip: Servo to angle
Knee: Passive swing
Ankle: Servo foot level to ground

Figure 4: Simulation Algorithm. Each leg has a state machine which is in one of four states. State transition conditions and actions in each state are shown.

to prevent it from banging into the knee cap and finally it is locked once it hits the knee cap.

In order for the natural swing property to exist on Spring Flamingo, we had to invert the direction of knee bend. Previously, the robot had a backward bent joint, meant to model the ankle of a bird. However, the robot did not have the joint corresponding to a birds knee. By giving the robot a human style leg configuration, it obtains passive swing characteristics.

In the next sections, we describe algorithms which exploit the three natural mechanisms, described above, in the control of a simulated and a physical bipedal robot.

3 Simulation Algorithm

We use the natural dynamic mechanisms described above in the control of a simulated seven link bipedal robot. The simulation has an actuated hip, knee, and ankle on each

Table 1: Physical parameters and controller parameters of the simulated planar bipedal walker. The controller parameters were tuned both manually and with a genetic algorithm.

Physical Parameter	Value
Total mass	14.2 kg
Body mass	12.0 kg
Hip to body center of mass	0.20 m
Body moment of inertia	0.10 kg m^2
Upper leg mass	0.46 kg
Upper leg moment of inertia	0.13 kg m^2
Upper leg length	0.42 m
Lower leg mass	0.31 kg
Lower leg moment of inertia	0.0095 kg m^2
Lower leg length	0.42 m
Foot mass	0.35 kg
Foot moment of inertia	0.0014 kg m^2
Foot height	0.04 m
Foot length forward	0.17
Foot length back	0.06

Controller Parameter	Manual	GA
hip swing angle	0.5 rad	0.56
hip gain	15.0	12.9
hip damp	1.0	1.76
hip hold angle	0.35	0.40
knee lock gain	1.0	0.68
knee damping	0.5	0.68
foot force threshold	20.0	22.5
swing time	0.4	0.45
ankle push gain	4.0	5.07
ankle compliance	1000	384
pitch gain	100	(same)
pitch damp	20	(same)
knee stance gain	30	(same)
knee stance damp	10	(same)

leg. It is confined to walk in the sagittal plane. The simulation parameters are listed in Table 1. Moments of inertia are measured about the center of mass. The center of mass of each leg link is approximately in the center of the link.

The simulation algorithm is summarized in Figure 4. Each leg acts separately and has a simple state machine. The leg can be in either Support, Toe Off, Swing, or Straighten states. In Support and Toe Off states, the hip is used to servo body pitch to maintain balance and the knee is locked to maintain height. In Support state, the ankle is unactuated - only the passive ankle compliance is present. During Toe Off state, the ankle is servoed to an angle using a Proportional-Derivative (PD) controller in addition to its passive compliance. The transition from Support to Toe Off occurs when the heel lifts off the ground due to the passive compliance of the ankle.

The robot transitions from Toe Off to Swing when the force on the foot falls below a certain threshold. In both Swing and Straighten states the hip is servoed to an angle using a PD controller and the foot is servoed to be level with the ground so that the robot does not stub its toe. In Swing state, the knee is damped while in Straighten state the knee is locked straight using a PD controller.

The robot transitions from Swing to Straighten state after a constant amount of time passes. Finally, the robot transitions from Straighten to Support state when the heel of the swing leg hits the ground.

The simulation parameters were first manually tuned, and then fine tuned using a genetic algorithm with efficiency as its cost function. Efficiency was computed as distance travelled divided by total joint energy after ten seconds of walking. Total joint energy was computed by integrating the total joint power which is the sum of the absolute values of the mechanical power at each joint:

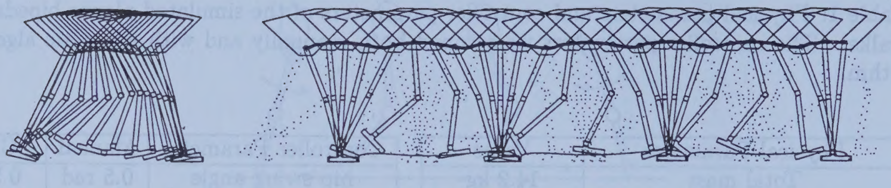


Figure 5: Elapsed time snapshot of the simulated robot walking data. The drawings on the left are spaced approximately 0.08 seconds apart and show the swing phase of one leg. The drawings on the right are spaced approximately 0.25 seconds apart. The right leg is dotted while the left leg is solid. Lines show the path of the tips of the feet and the hip trajectory. The robot walks from left to right.

$$E_{tot} = \int P_{tot}\, dt, \quad P_{tot} = \sum_{joints} |P_{joint}|, \quad P_{joint} = \tau_{joint}\, \dot{\theta}_{joint} \tag{1}$$

After a couple generations, naturally looking simulations resulted. Some of them achieved a smooth transfer of support by coordinating toe off and heel strike. One of those solutions was chosen to be presented here as it appeared quite natural. A time elapsed animation is shown in Figure 5. The drawings on the left show the swing phase of one leg. The drawings on the right show several steps. The right leg is dotted while the left leg is solid. Lines show the path of the tips of the feet and the hip trajectory. Several variables are plotted graphically in Figure 6. In the first row, left to right, are the state of the left leg (solid) and right leg (dotted), forward distance travelled (x), and velocity. In the second row are total power, body height, and pitch. In the last row are hip, knee, and ankle power of the left leg.

We see that the simulated robot walked at a moderate speed (approximately 0.8 m/s) and had a natural looking gait. The simulation requires less than 20 watts of joint power on average. Most of the energy consumption comes from the hips at the beginning of swing phase. It is interesting that the algorithm does not contain any explicit speed control mechanism, yet speed is stabilized. We speculate that this is due to the natural system dynamics, in the same way that speed is naturally stabilized in the passive dynamic walkers.

4 Robot Algorithm

Figure 7 contains photographs of Spring Flamingo, a seven link planar bipedal walking robot. The photo on the left is a close up of the knee joint and kneecap. The photo on the right is a close up of the compliant ankle stop. We exploit these two natural mechanisms, and the passive swing in the control of the robot.

The robot algorithm is summarized in Figure 8. As in the simulation algorithm, each leg acts separately and has a simple state machine. The robot algorithm differs from the simulation algorithm in two ways. First, because the robot does not have ground force sensors yet, we use geometric state transition conditions for initiation of swing. Second, instead of controlling each joint separately, in a decoupled manner, we use Virtual Model Control [14, 15] in the control of pitch and height.

The control parameters for the physical robot were manually tuned over the course of two days. Stable solutions were found that were fairly robust such that the robot

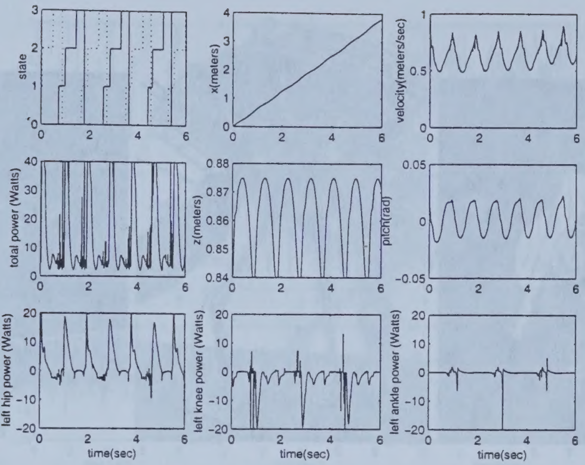


Figure 6: Simulation data. The first row contains, left to right, state of the legs, forward distance travelled, and forward velocity. The second row contains total power, body height, and body pitch. The last row contains joint power at the hip, knee, and ankle of the left leg.

walked continuously on level ground. A time elapsed animation of the robot is shown in Figure 9. The drawings on the left show the swing phase of one leg. The drawings on the right show several steps. Several variables are plotted graphically in Figure 10. In the first row, left to right, are the state of the left leg (solid) and right leg (dotted), forward distance travelled (x), and velocity. In the second row are total power, body height, and pitch. In the last row are hip, knee, and ankle power of the left leg.

We see that the physical robot walked at a casual speed (approximately 0.5 m/s) and had a natural looking gait. The robot requires less than 10 watts of joint power on average. As in the simulation, most of the energy consumption comes from the hips at the beginning of swing phase. Again, as in the simulation, it is interesting that the algorithm does not contain any explicit speed control mechanism, yet speed is stabilized.

5 Conclusions and Future Work

Spring Flamingo presently walks continuously by exploiting the natural dynamics of a kneecap, compliant ankle, and passive swing leg. Two very simple algorithms are used in simulation and on the physical robot. The resultant motion is fairly smooth and efficient. This work may help bridge the gap between passive dynamic walkers and powered bipedal robots.

Both the simulation and the robot settle on a stable speed of walking. The simulation walks approximately 0.8 m/s while the robot walks approximately 0.5 m/s. However, nowhere in either controller is speed explicitly controlled. We believe that the speed is stabilized in a similar way to passive dynamic walking machines. That is, if the robot goes too fast, it naturally takes a longer step due to the natural swing leg dynamics and hence slows down on the next step. Similarly, if the robot moves too slowly, it naturally takes a shorter step and hence speeds up on the next step. Speed is dependent

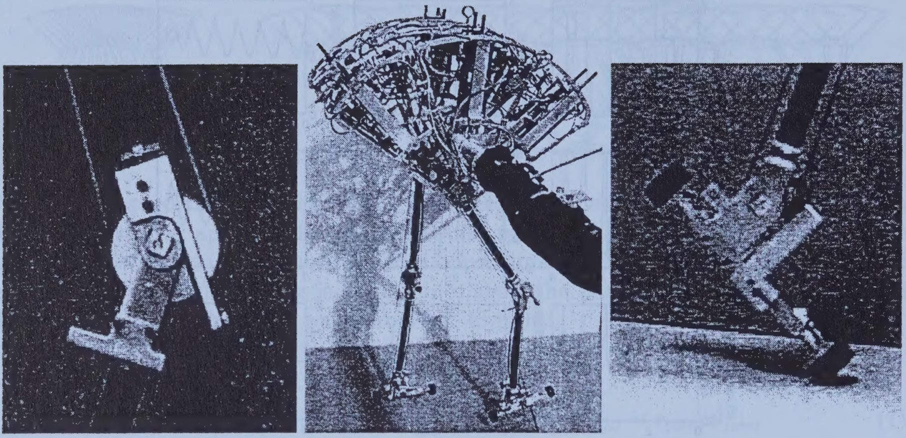


Figure 7: Spring Flamingo, a planar bipedal walking robot. There are six force controlled actuators attached to the body. Power is transmitted to the hips, knees, and ankles via cables. A boom prevents motion in the lateral, roll, and yaw directions.

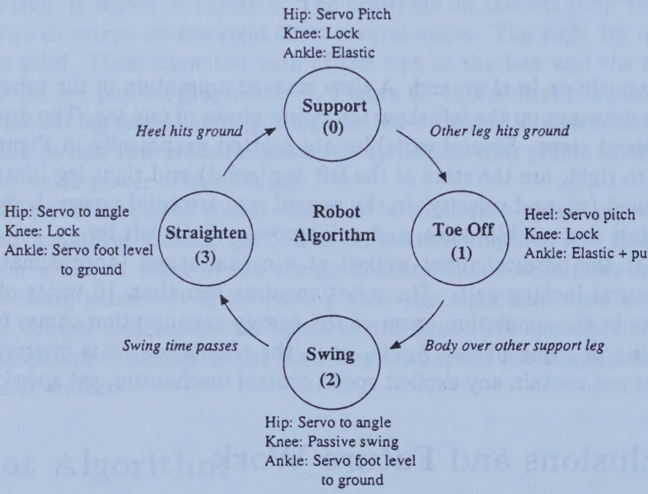

Hip: Servo Pitch
Knee: Lock
Ankle: Elastic
Support
(0)
Heel hits ground
Other leg hits ground
Hip: Servo to angle
Knee: Lock
Ankle: Servo foot level to ground
Straighten
(3)
Robot Algorithm
Toe Off
(1)
Heel: Servo pitch
Knee: Lock
Ankle: Elastic + push
Swing time passes
Body over other support leg
Swing
(2)
Hip: Servo to angle
Knee: Passive swing
Ankle: Servo foot level to ground

Figure 8: Physical Robot Algorithm. Each leg has a state machine which is in one of four states. State transition conditions and actions in each state are shown.

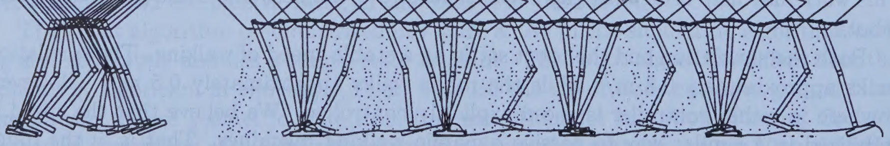


Figure 9: Elapsed time snapshot of the physical robot walking data. The drawings on the left are spaced approximately 0.07 seconds apart and show the swing phase of one leg. The drawings on the right are spaced approximately 0.4 seconds apart.

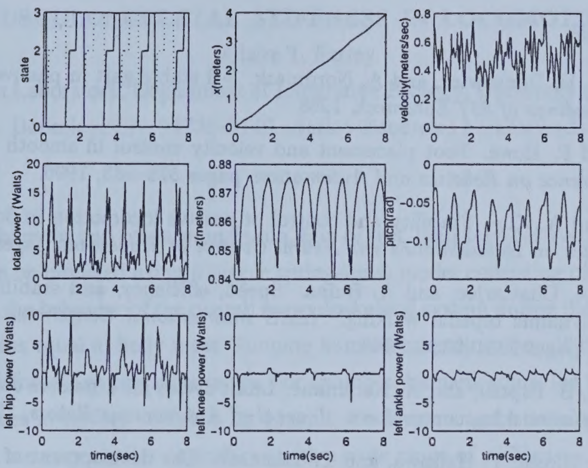


Figure 10: Physical robot walking data.

on parameter values. We plan on controlling speed and achieving faster walking through the modulation of these parameter values.

In order to exploit the natural dynamics of a walking robot, it is important that the inertia and friction of the actuators does not dominate the dynamics of the legs. On Spring Flamingo we use Series Elastic Actuators which result in very little unwanted dynamics. These actuators have good force dynamic range and low force offset which are important in natural and efficient walking. We are continuing to use these actuators in the design of several new walking robots.

Both the simulation and the physical robot required tuning of parameters in order to walk. The simulation was tuned first manually and then fine tuned using a genetic algorithm with efficiency as the cost function. The genetic algorithm found a set of parameters which nicely synchronized toe off of the support leg with heel strike of the swing leg such that a smooth transfer of support occurred and minimal energy was lost. The robot was only tuned manually but still walked fairly well. Its joint power averaged less than 10 watts. We are also considering tuning the robot using a genetic algorithm or other automatic techniques.

Spring Flamingo demonstrates that natural mechanisms can be exploited to make control of bipedal robots simple and the resultant motion natural looking and efficient. With minor changes to the algorithm, the robot should have the capability to start and stop and traverse slopes and go over obstacles. These results give us confidence that we can develop similar algorithms for three dimensional bipedal walking.

Acknowledgements

Dan Paluska designed and built the compliant ankle for Spring Flamingo. Dave Robinson designed and built the feet and force sensors. Spring Flamingo's actuators are based on a design by Mike Wittig. Dan, Dave, and Mike helped in many other aspects of the robot's design and construction.

References

[1] J. Adolfsson, H. Dankowicz, and A. Nordmark. 3-d stable gait in passive bipedal mechanisms. *Proceedings of 357 Euromech*, 1998.

[2] E. Dunn and R. Howe. Foot placement and velocity control in smooth bipedal walking. *IEEE Conference on Robotics and Automation*, pages 578–583, 1996.

[3] J. Fowble and A. Kuo. Stability and control of passive locomotion in 3d. *Proceedings of the Conference on Biomechanics and Neural Control of Movement*, pages 28–29, 1996.

[4] M. Garcia, A. Chatterjee, and A. Ruina. Speed, efficiency, and stability of small-slope 2d passive dynamic bipedal walking. *IEEE International Conference on Robotics and Automation*, pages 2351–2356, 1998.

[5] A. Goswami, B. Espiau, and A. Keramane. Limit cycles ini a passive compass gait biped and passivity-mimicking control laws. *Journal of Autonomous Robots*, 1997.

[6] K. Hirai, M. Hirose, Y. Haikawa, and T. Takenaka. The development of honda humanoid robot. *IEEE Conference on Robotics and Automation*, 1998.

[7] L. Jalics, H. Hemami, and B. Clymer. A control strategy for adaptive bipedal locomotion. *IEEE Conference on Robotics and Automation*, pages 563–569, 1996.

[8] S. Kajita and K.Tani. Adaptive gait control of a biped robot based on realtime sensing of the ground profile. *IEEE Conference on Robotics and Automation*, pages 570–577, 1996.

[9] A. Kun and W. T. Miller. Adaptive dynamic balance of a biped robot using neural networks. *IEEE Conference on Robotics and Automation*, pages 240–245, 1996.

[10] Tad McGeer. Passive dynamic walking. *International Journal of Robotics Research*, 9(2):62–82, 1990.

[11] H. Miura and I. Shimoyama. Dynamic walk of a biped. *International Journal of Robotics Research*, 3(2):60–74, 1984.

[12] Simon Mochon and Thomas A. McMahon. Ballistic walking: An improved model. *Mathematical Biosciences*, 52:241–260, 1979.

[13] Gill A. Pratt and Matthew M. Williamson. Series elastic actuators. *IEEE International Conference on Intelligent Robots and Systems*, 1:399–406, 1995.

[14] J. Pratt, P. Dilworth, and G. Pratt. Virtual model control of a bipedal walking robot. *IEEE Conference on Robotics and Automation*, pages 193–198, 1997.

[15] J. Pratt and G. Pratt. Intuitive control of a planar bipedal walking robot. *IEEE Conference on Robotics and Automation*, 1998.

[16] Marc H. Raibert. *Legged Robots That Balance.* MIT Press, Cambridge, MA, 1986.

[17] M. Vukobratovic, B. Borovac, D. Surla, and D. Stokic. *Biped Locomotion: Dynamics, Stability, Control, and Applications.* Springer-Verlag, Berlin, 1990.

[18] J. Yamaguchi, A. Takanishi, and I. Kato. Development of a biped walking robot adapting to a horizontally uneven surface. *IEEE International Conference on Intelligent Robots and Systems*, pages 1156–1163, 1994.

[19] K. Yi and Y. Zheng. Biped locomotion by reduced ankle power. *IEEE Conference on Robotics and Automation*, pages 584–589, 1996.

MUSCULOSKELETAL STIFFNESS IN LOCOMOTION

Claire T. Farley
Locomotion Laboratory, Department of Integrative Biology, University of California,
Berkeley, CA. 94720-3140. cfarley@socrates.berkeley.edu

Abstract

When humans and other animals run, hop or trot, their legs behave like springs. These gaits can be modeled using a simple spring-mass model consisting of a "leg spring" that represents the behavior of the overall musculoskeletal system during the ground contact phase and a mass equal to body mass. Running humans can adjust the stiffness of their leg spring ("leg stiffness") over a large range. For example, humans adjust leg stiffness to offset differences in surface stiffness. As a result, the dynamics of a running human's center of mass are very similar on surfaces with a wide range of stiffnesses. We have examined humans hopping in place in order to uncover the mechanisms used to adjust leg stiffness. Leg stiffness is very sensitive to changes in ankle torsional stiffness but relatively insensitive to changes in knee or hip stiffness during hopping in place. Furthermore, experiments have revealed that hopping humans adjust leg stiffness primarily by changing ankle stiffness.

Spring-mass model

Running, hopping and trotting animals all bounce along the ground in a spring-like manner (4, 6, 12, 16, 20). During these gaits, the actions of the body's many musculoskeletal elements, including muscles, tendons and ligaments, are integrated together so that the overall musculoskeletal system behaves like a single spring. Consequently, these gaits can be modeled using a simple spring-mass system, consisting of a single linear "leg spring" and a point mass equal to body mass (Figure 1) (2, 3, 5, 8-10, 14, 15, 19, 22). In this model, the leg spring compresses and then lengthens during the ground contact phase. Similarly, in a running animal, the leg shortens as limb joints flex during the first half of the ground contact phase and then lengthens as the limb joints extend during the second half of the ground contact phase. In spite of the simplicity of the spring-mass model relative to the complexity of actual neuromuscular system, it describes the mechanics of bouncing gaits remarkably well.

The stiffness of the leg spring represents the average stiffness of the overall musculoskeletal system during the ground contact phase ("leg stiffness"). Leg stiffness influences the mechanics and kinematics of the body's interaction with the ground. For example, a greater leg stiffness leads to a shorter time of foot-ground contact and a smaller vertical excursion of the body's center of mass during the ground contact phase.

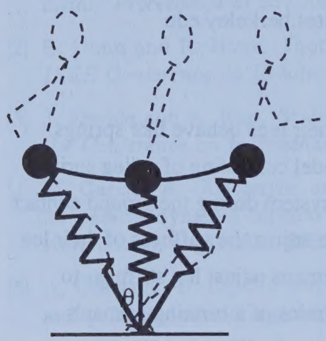


Figure 1: The spring-mass model consists of a leg spring and a mass. A stick figure of a runner is shown in dashed lines. For both the model and the runner, a single ground contact phase is shown. From force platform and kinematic data, leg stiffness can be calculated from the relationship between the force in the leg spring (i.e., ground reaction force) and leg spring compression (i.e., change in leg spring length relative to the beginning of the ground contact phase).

Adjustments for speed and stride frequency in running humans and robots

Leg stiffness remains the same at all forward speeds in running humans (19). Humans are able to run at higher speeds, and with shorter ground contact times, by increasing the angle swept by the leg during the stance phase. A variety of hopping, trotting, and running animals keep leg stiffness the same at all speeds and alter the angle swept by the leg to adjust for different speeds (9). In all of these animals, the leg sweeps through a larger angle at higher speeds, thus reducing the vertical movements of the center of mass at higher speeds.

It is possible for the stiffness of the leg to be adjusted during bouncing gaits. When humans hop in place, a very simple bouncing gait, leg stiffness is increased in order to increase hopping frequency or hopping height (8). Similarly, when humans run, the stiffness of the leg is adjusted to allow the use of a range of stride frequencies at a given speed (10). A stiffer leg leads to a higher stride frequency and shorter stride length at a given speed. Thus, during overground running in the natural world, it is likely that adjustments to leg stiffness allow runners to alter stride length to avoid obstacles.

The fastest robots have spring-based legs (25-28). We can gain insight into the control strategies that are feasible for animal locomotion by exploring the range of workable strategies in these robots. These robots use either compressed air or metal springs in their legs to store and return elastic energy with each step as they run, hop, or trot. There are many similarities between running in humans and bouncing gaits in these robots. Both these robots and animals run at different speeds by altering the angle swept by their legs while keeping their leg stiffness the same (9, 19, 28). In addition, altering the robot's leg stiffness leads to changes in stride frequency at the same speed (28), just as observed in humans (8, 10). The control of these robots is greatly simplified by relying on the passive dynamics of spring-mass system of the robot's body. The robot's movements are largely determined by physical parameters including the stiffness of the leg spring, the angle at which the leg spring is set down, and the mass of the robot (25). The control algorithms work in concert with the physical properties of the robot's body to produce stable locomotion. It seems logical to suggest that animals rely on the spring-mass dynamics of their bodies in a similar manner, thus simplifying the neural control of locomotion.

Adjustments for different surfaces

Recent studies have revealed that humans adjust leg stiffness to accommodate elastic surfaces of different stiffnesses (14, 15). For example, when humans hop in place, leg stiffness is increased by as much as 3-fold to accommodate reductions in surface stiffness (Figure 2A; 14). Similarly, when humans run at a given speed, leg stiffness is adjusted to offset changes in surface stiffness (15). During both hopping in place and forward running, the adjustment to leg stiffness allows the total stiffness of the series combination of the leg and the surface to remain the same on all surfaces (14, 15). As a result, the ground contact time and the vertical displacement of the center of mass during ground contact remain the same on surfaces with a 1000-fold range of stiffnesses. If leg stiffness were not adjusted to accommodate decreases in surface stiffness, the feet would remain in contact with the ground for a longer time, and the center of mass would move a larger vertical distance while the feet were on the ground. Thus, the adjustments to leg stiffness are critical to allowing similar locomotion on surfaces of different stiffnesses.

Potential mechanisms for leg stiffness adjustment

It is important to realize that the spring-mass model represents the behavior, but not the structure, of the integrated musculoskeletal system during bouncing gaits. The compression and extension of the leg spring that occurs during the ground contact phase corresponds to joints flexing and extending. For example, during hopping in place, the ankle, knee, and hip flex during the first half of the ground contact phase and extend during the second half of the ground contact phase. In a multi-jointed system like the human musculoskeletal system, the overall stiffness (i.e., the leg stiffness) depends on a combination of the torsional stiffnesses of the joints and the geometry of the system. The torsional stiffness of each joint determines how much its angle changes in response to a given moment about the joint. If the ankle, knee, and hip are stiffer, they will undergo smaller angular displacements during the ground contact phase, resulting in less leg compression and a higher leg stiffness. Many studies have demonstrated that joint stiffness is highly adjustable during single joint movements. For single joint movements involving the ankle, joint stiffness depends on many factors including muscle activation, reflexes, joint moment, and joint angle (1, 17, 21, 24, 29-32). Thus, it is possible that the mechanism for adjusting leg stiffness involves adjusting the stiffness of the leg joints.

Leg stiffness during locomotion is also influenced by the geometry of the leg at the instant that the foot hits the ground ("touchdown leg geometry") (18, 23). Mathematical models predict that leg stiffness should depend on touchdown leg geometry due to changes in the alignment of the ground reaction force vector relative to the joints (18, 23). If the leg joints are more flexed when the foot hits the ground, the joint moments associated with a given ground reaction force will be greater. Thus, for a given joint stiffness, the angular displacement of the joints during the ground contact phase also will be greater. These mathematical models have focused on how leg geometry influences leg stiffness in the absence of changes to joint stiffness. In the real musculoskeletal system, however, leg geometry can also influence joint stiffness because it affects muscle-tendon length and the level of muscle activation required to apply a given force to the ground (1, 17, 21, 24, 29-32).

Direct experiments on humans have demonstrated ("Groucho running", 23), leg stiffness is substantially lower than during normal running. Similarly, when humans bounce on a compliant board, adopting a posture with greater knee flexion leads to a lower leg stiffness (18). Finally, when humans land from a jump, the stiffness of the landing appears to depend on knee angle at landing (7). In all of these activities, it is clear that changes in leg geometry affect the alignment of the ground reaction force vector. However, there is no information about whether the changes in leg geometry affect joint stiffness due to changes in the lengths of muscle-tendon units crossing the leg joints.

Mechanisms for leg stiffness adjustment: insights from hopping in place

We have examined the mechanisms used by humans hopping in place to adjust the leg stiffness. When humans hop on surfaces of different stiffnesses, they increase leg stiffness to offset decreases in surface stiffness. Humans increase leg stiffness primarily by increasing the torsional stiffness of the ankle. In response to decreases in surface stiffness, ankle stiffness increases by about 1.75-fold (Figure 2B; 11). Furthermore, when humans increase hopping height while hopping at a given frequency, leg stiffness increases by as much as 2-fold while ankle stiffness also increases by nearly 2-fold (13).

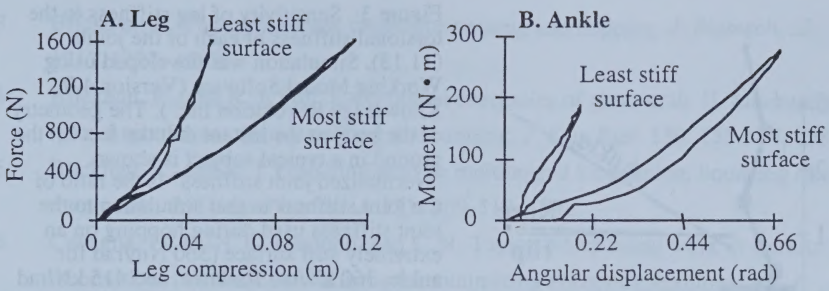


Figure 2. (A) Force in the leg spring versus leg spring compression for a single ground contact phase of hopping on an extremely stiff surface (35,000 kN/m) and on a low stiffness surface (22 kN/m). The force begins to increase and the leg spring begins to compress when the feet hit the ground. The force and leg compression are maximum at the middle of the ground contact phase. Average leg stiffness is calculated by taking the ratio of the ground reaction force to leg compression at the middle of the ground contact phase. **(B)** Ankle net muscle moment (i.e., net moment generated by all muscles crossing the ankle) versus angular displacement for a single ground contact phase on both surfaces. The ankle moment increases above zero (i.e., extensor moment) as soon as the feet hit the ground. Simultaneously, the angle between the foot and the shank begins to decrease, and thus, ankle "angular displacement" relative to the instant of touchdown begins to increase. At the middle of the ground contact phase, both net muscle moment and angular displacement reach their maximum values. The average ankle stiffness is calculated from the ratio of the moment to the angular displacement at the middle of the ground contact phase.

Using a computer simulation of a human hopping in place, we examined the effects of adjusting ankle stiffness on leg stiffness (11, 13). Our computer model consists of four segments (foot, shank, thigh, trunk) connected by three torsional springs representing the joints (ankle, knee, hip). The segment properties, segment initial conditions, and the torsional spring properties are identical to those of a typical subject at the instant that the feet hit the ground. The simulation is allowed to run, and the model interacts with the ground. The leg stiffness of the model is calculated in exactly the same way as in human hoppers (see Figure 2 legend). Sensitivity analyses revealed that the model's leg stiffness is most sensitive to changes in ankle torsional stiffness and much less sensitive to changes in knee or hip stiffness (Figure 3). Leg stiffness increases by nearly 2-fold when ankle stiffness doubles but only increases by 1.08-fold when knee stiffness doubles and 1.03-fold when hip stiffness doubles (11).

Controlling the stiffness of the human leg during hopping in place is much simpler than we expected. The nervous system can control the stiffness of the overall leg simply by controlling the torsional stiffness of a single joint, the ankle. It seems likely that the knee and hip stiffness values are important in the control of fore-aft stability in a human hopper although no studies have directly examined this issue.

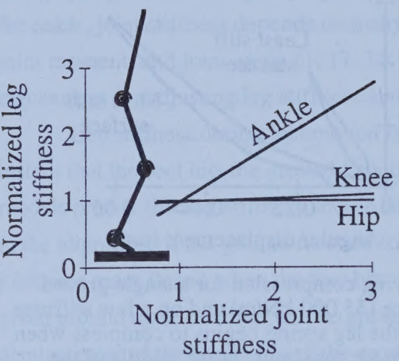


Figure 3: Sensitivity of leg stiffness to the torsional stiffness of each of the joints (11,13). Simulation was developed using Working Model Software (Version 4.0, Knowledge Revolution Inc.). The geometry of the body at the instant that the feet hit the ground in a typical subject is shown. "Normalized joint stiffness" is the ratio of the joint stiffness in that simulation to the joint stiffness used during hopping on an extremely stiff surface (360 Nm/rad for ankle, 360 kN/rad for knee, and 415 kN/rad for hip). "Normalized leg stiffness" is the ratio of the leg stiffness in that simulation to the leg stiffness used during normal hopping on an extremely stiff surface. The data for each joint were obtained by keeping the stiffness of the other two joints and leg geometry at touchdown the same as on the most stiff surface, and then systematically changing the stiffness of the joint of interest.

Conclusions

The overall stiffness of the leg is extremely important in determining many aspects of bouncing gaits. When a human runs or hops in place, leg stiffness is a key determinant of the time of foot-ground contact and the trajectory of the center of mass. Leg stiffness can be adjusted to allow running or hopping on different surfaces or at different stride frequencies. When humans hopping in place, leg stiffness primarily depends on the stiffness of the ankle. Future research should examine whether this selective sensitivity of leg stiffness to ankle stiffness extends to running, and should examine the neuromuscular determinants of ankle stiffness.

Acknowledgments

This research was supported by a National Institutes of Health grant (R29 AR44008).

References

1. Agarwal, G. C., and G. L. Gottlieb. Oscillation of the human ankle joint in response to applied sinusoidal torque on the foot. *J. Physiol. (Lond.)* 268: 151-176, 1977.
2. Alexander, R. M. *Elastic Mechanisms in Animal Movement.* Cambridge: Cambridge University Press, 1988.

3. Blickhan, R. The spring-mass model for running and hopping. *J. Biomech.* 22: 1217-1227, 1989.
4. Blickhan, R., and R. J. Full. Locomotion energetics of ghost crab. II. Mechanics of the center of mass during walking and running. *J. Exp. Biol.* 130: 155-174, 1987.
5. Blickhan, R., and R. J. Full. Similarity in multilegged locomotion: bouncing like a monopode. *J. Comp. Physiol. A* 173: 509-517, 1993.
6. Cavagna, G. A., N. C. Heglund, and C. R. Taylor. Mechanical work in terrestrial locomotion: two basic mechanisms for minimizing energy expenditure. *Am. J. Physiol.* 233: R243-R261, 1977.
7. Devita, P., and W. A. Skelly. Effect of landing stiffness on joint kinetics and energetics in the lower extremity. *Med. Sci. Sports Exerc.* 24: 108-115, 1992.
8. Farley, C. T., R. Blickhan, J. Saito, and C. R. Taylor. Hopping frequency in humans: a test of how springs set stride frequency in bouncing gaits. *J. Appl. Physiol.* 71: 2127-2132, 1991.
9. Farley, C. T., J. Glasheen, and T. A. McMahon. Running springs: speed and animal size. *J. Exp. Biol.* 185: 71-86, 1993.
10. Farley, C. T., and O. Gonzalez. Leg stiffness and stride frequency in human running. *J. Biomech.* 29: 181-186, 1996.
11. Farley, C. T., H. H. P. Houdijk, C. van Strien, and M. Louie. Mechanism of leg stiffness adjustment for hopping on surfaces of different stiffnesses. *J. Appl. Physiol.* 85: 1044-1055, 1998.
12. Farley, C. T., and T. C. Ko. Mechanics of locomotion in lizards. *J. Exp. Biol.* 200: 2177-2188, 1997.
13. Farley, C. T., and D. C. Morgenroth. Leg stiffness primarily depends on ankle stiffness during human hopping. *J. Biomech.* , in press.
14. Ferris, D. P., and C. T. Farley. Interaction of leg stiffness and surface stiffness during human hopping. *J. Appl. Physiol.* 82: 15-22, 1997.
15. Ferris, D. P., M. Louie, and C. T. Farley. Running in the real world: adjusting leg stiffness for different surfaces. *Proc. Roy. Soc. Lond. B* 265: 989-994, 1998.
16. Full, R. J., and M. S. Tu. Mechanics of six-legged runners. *J. Exp. Biol.* 148: 129-146, 1990.
17. Gottlieb, G. L., and G. C. Agarwal. Dependence of human ankle compliance on joint angle. *J. Biomech.* 11: 177-181., 1978.
18. Greene, P. R., and T. A. McMahon. Reflex stiffness of man's anti-gravity muscles during kneebends while carrying extra weights. *J. Biomech.* 12: 881-91, 1979.

19. He, J. P., R. Kram, and T. A. McMahon. Mechanics of running under simulated low gravity. *J. Appl. Physiol.* 71: 863-870, 1991.
20. Heglund, N. C., G. A. Cavagna, and C. R. Taylor. Energetics and mechanics of terrestrial locomotion. III. Energy changes of the centre of mass as a function of speed and body size in birds and mammals. *J. Exp. Biol.* 97: 41-56, 1982.
21. Hunter, I. W., and R. E. Kearney. Dynamics of human ankle stiffness: variation with mean ankle torque. *J. Biomech.* 15: 747-52, 1982.
22. McMahon, T. A., and G. C. Cheng. The mechanics of running: how does stiffness couple with speed? *J. Biomech.* 23 (suppl. 1): 65-78, 1990.
23. McMahon, T. A., G. Valiant, and E. C. Frederick. Groucho running. *J. Appl. Physiol.* 62: 2326-2337, 1987.
24. Nielsen, J., T. Sinkjaer, E. Toft, and Y. Kagamihara. Segmental reflexes and ankle joint stiffness during co-contraction of antagonistic ankle muscles in man. *Exp. Brain Res.* 102: 350-8, 1994.
25. Raibert, M. H. *Legged Robots That Balance*. Cambridge, MA: MIT Press, 1986.
26. Raibert, M. H. Trotting, pacing and bounding by a quadruped robot. *J. Biomech.* 23 (suppl. 1): 79-98, 1990.
27. Raibert, M. H., and J. K. Hodgins. Legged robots. In: *Biological Neural Networks in Neuroethology and Robotics*, edited by R. D. Beer, R. E. Ritzmann and T. McKenna. Boston: Academic Press, 1993, p. 319-354.
28. Raibert, M. H., R. R. Playter, J. K. Hodgins, R. R. Ringrose, C. Francois, Y. Bendana, G. Zeglin, and L. Borvansky. Dynamically stable legged locomotion. . Cambridge, MA: Massachusetts Institute of Technology, 1993.
29. Sinkjaer, T., E. Toft, S. Andreassen, and B. C. Hornemann. Muscle stiffness in human ankle dorsiflexors: intrinsic and reflex components. *J. Neurophysiol.* 60: 1110-21, 1988.
30. Weiss, P. L., I. W. Hunter, and R. E. Kearney. Human ankle joint stiffness over the full range of muscle activation levels. *J. Biomech.* 21: 539-544, 1988.
31. Weiss, P. L., R. E. Kearney, and I. W. Hunter. Position dependence of ankle joint dynamics--I. Passive mechanics. *J Biomech* 19: 727-35, 1986.
32. Weiss, P. L., R. E. Kearney, and I. W. Hunter. Position dependence of ankle joint dynamics--II. Active mechanics. *J Biomech* 19: 737-51, 1986.

Metabolic Energy Consumption During Animal Locomotion

Rodger Kram
Locomotion Lab
Integrative Biology Dept.
University of California
3060 VLSB
Berkeley, CA 94720-3140
rkram@socrates.berkeley.edu
http://socrates.berkeley.edu/~rkram/

Abstract

How do animals produce smooth locomotor movements ? One possibility is that animals sense the metabolic energy consumed and adjust their movements to minimize metabolic energy consumption. An analogous control scheme for locomoting robots might provide advantages. Most engineers would assume that to minimize metabolic energy consumption would simply require a minimization of the mechanical work performed by the actuators. However, in biological animals, a very large portion of the energy consumed by the "actuators", i.e. the muscles, is for the production of nearly isometric force with no length change and thus little mechanical work. Phenomenologically, there is much evidence that animals do tune their locomotion to minimize metabolic energy consumption. In most situations, humans and other animals choose the energetically optimal gaits, speeds and stride frequencies. Animals have sensors for many variables related to metabolic energy consumption such as muscle-tendon force, limb joint excursion, heart rate and breathing frequency. However, there are no direct sensors of metabolic energy consumption.

Introduction

Muscles are the actuators for animal movement and they are the major consumer of metabolic energy during locomotion. In this paper, I weight the evidence for the idea that animal locomotor movements are optimized to minimize metabolic energy consumption. I begin with a discussion of the most relevant performance characteristics of muscle, e.g. force, velocity and power. I then explain our current understanding of the link between the mechanics of locomotion and the metabolic cost of locomotion. Next, I present evidence that animals optimize the speed, gait and kinematics that minimize metabolic energy consumption. Finally, I consider if animals have adequate biological sensors and "software" to determine their rate of metabolic energy consumption.

Muscle

Muscles exert force by the "rowing" of small molecular oars (called cross bridges) made of protein and powered by the nearly universal biological fuel source ATP (adenosine triphosphate). Muscles exert only tension forces, but in locomotion those tension forces may involve muscle shortening, muscle force with no lengthening (isometric) or provide active resistance to lengthening. When a muscle is fully activated but not allowed to shorten, it can exert a tension stress of ~ 25 N/cm^2, that is, force is proportional to cross sectional area of muscle fibers activated. The maximal force per cross sectional area does not vary substantially from muscle to muscle nor across the variety of vertebrate animals. The maximal force exerted by a muscle does not depend on the length of the muscle fibers themselves, although when a muscle is at short or long relative lengths, it is substantially weaker. This is because the molecular filaments slide relative to each other such that fewer cross bridges can attach to binding sites. Usually muscles in animals function over lengths where force does not vary.

The force exerted by a muscle depends on the velocity of shortening. A great deal of force can be produced isometrically, that is when length does not change. Muscles produce proportionately less force at faster velocities. If a muscle is neurally stimulated without any load it shortens at maximal velocity (v_{max}). V_{max} varies dramatically from muscle fiber to muscle fiber, from muscle to muscle and from animal to animal. Eye blink muscles are faster than leg muscles and a leg muscle of a mouse is faster than a leg muscle of an elephant. Muscles provide very large forces when they resist being lengthened. Because force varies with velocity, the product of force and velocity, or power production also varies with velocity.

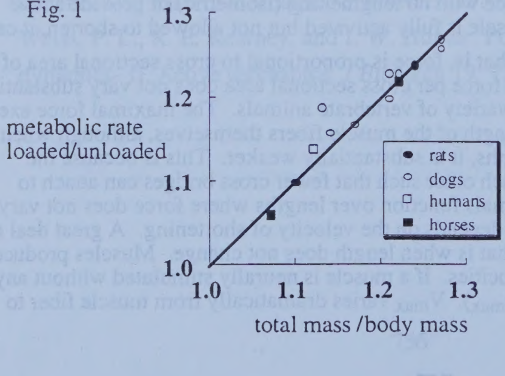
The production of muscle force begins with neural activation. The rate of force rise depends on the stimulation, but more importantly here, on the intrinsic rate at which the crossbridges (protein oars) cycle through their rowing movements. That intrinsic rate is set by the biochemical catalysts or enzymes (especially myosin ATP-ase) and the molecular structure of the crossbridges. It is believed that each cycle of each cross bridge consumes one molecule of ATP, the universal biological energy source. Thus, faster rates of cross bridge cycling involve faster rates of metabolic energy consumption. To exert greater forces, more muscle fibers must be activated in parallel , and thus energy consumption is also proportional to force. If a muscle fiber is longer, more crossbridges are involved and again, the rate of energy consumption is greater. In short, the rate of metabolic energy consumption is proportional to the cross sectional area, fiber length and intrinsic cross bridge cycling rate. If a muscle exerts force *and* shortens and performs work, then even more metabolic energy is consumed. Tom McMahon's book (11) provides an excellent introduction to these biological processes and is written by and for engineers.

Mechanical-Metabolic Link

What mechanical processes determine the rate of energy consumption of locomoting animals? The muscles of walking and running animals must perform three mechanical tasks, each of which consumes some metabolic energy: 1. The muscles must swing the legs forward for each stride. 2. The muscles must shorten and perform mechanical work to lift and lower, accelerate and decelerate the center of mass. 3. The muscles must generate force to prevent the legs from buckling, i.e. they must support the weight of the body against gravity.

Experimental evidence strongly suggests that the metabolic cost for swinging the legs relative to the center of mass of the whole body is rather small. Taylor et al. (16) performed an elegant experiment on a range of small to large species of mammals (rats, dogs, humans and horses). They trained the animals to run on a treadmill while wearing added mass in vests attached firmly to their torsos. While the animals ran, the scientists measured their rates of oxygen consumption which is our best measure of the metabolic energy consumed. The loads were carried near the animal's center of mass and not on the limbs. Stride frequency and other limb kinematics were unchanged. If the metabolic cost of swinging the legs was substantial say, 50% of the total during normal unloaded running, then presumably Taylor et al. would have found only a 5% increase in metabolic rate when a 10% load was carried. However, that was not the case. They found that at a given running speed, the rate of energy consumption increased in direct proportion to the load carried when expressed as a per cent of original body mass. That is, a 10% of body mass load increased the metabolic rate 10% and like wise a 20% load increased metabolic rate 20% and so on. See Figure 1. From this they inferred that the metabolic cost of leg swinging is minimal.

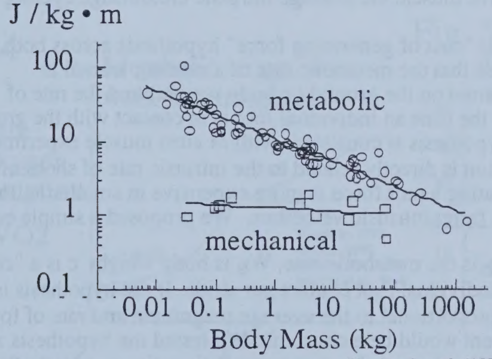
Fig. 1



Locomotion is metabolically much more expensive per unit body mass for smaller animals. A mouse consumes more than 10 times as much energy to transport a gram of it's body a meter forward compared to a horse. One explanation might be that mouse muscles are somehow less efficient than horses. However, evidence from muscles extracted from animals and studied in controlled physiological solutions suggests that this is not the case. All vertebrate muscles perform work with nearly the same efficiency, that is the metabolic cost per unit of mechanical work is essentially constant. Given this, the high cost of locomotion by small animals could be explained if they must perform more mechanical work.

A comprehensive comparative four part study of both the metabolic cost of locomotion and the mechanical work performed during locomotion was conducted under the leadership of C. Richard Taylor in 1982 (4-6, 15). More than 20 different species of birds and mammals were studied ranging from 30 gram birds to 254 kg Zebu cattle. The animals all ran on treadmills while wearing oxygen consumption measuring masks. This allowed the physiologists to determine the metabolic cost of transport, that is the Joules of metabolic energy consumed in moving a kilogram of animal one meter forward. Simultaneously, high speed films were taken to establish the movements of the limbs. Finally, a subset of the animals ran across force platforms so as to record the mechanical work performed on the center of mass. Using conservative assumptions about elastic energy storage and recovery as well as intersegmental transfer of mechanical energy, they obtained measures for the mechanical work of transport, that is the Joules of work performed to move a kilogram of animal mass a meter forward. The overriding finding of this tour de force was that while smaller animals have a much greater metabolic cost of transport compared to large animals, they all performed nearly the same amount of mechanical work per kilogram per meter (see Fig. 2). At first, this seems in conflict with the fact that isolated muscles from these animals are equally efficient. However, it is important to understand that walking and running animals have remarkably effective mechanisms for minimizing the amount of work that the muscles must perform.

Fig. 2



Heglund et al. (1982)

Walking involves an inverted pendulum-like movement of the body. A useful mechanical model of walking simplifies the upper body to a point mass located at the hip and the legs to be rigid struts. In this model, the path of the body follows a series of arcs as it vaults over the supporting leg. During the first half of this arc, the body slows in the forward direction (losing kinetic energy) but rises and gains gravitational potential energy. During the second half of the stance phase the gravitational potential energy is transferred back into forward kinetic energy. In reality, this pendulum-like exchange of energy is not perfect, about 65% of the energy is recovered with each step and so the muscles must shorten and perform work to make up for the loss of energy. Throughout the stance phase of walking, the knee extensor muscles in particular must be active to allow the leg to act like a rigid strut. Thus, walking involves some muscle activity to perform work on the center of mass and substantial muscle activity to support body weight.

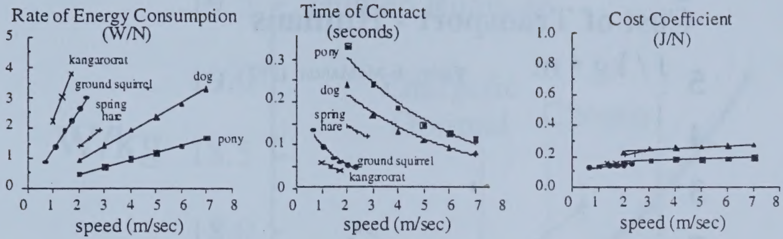
In contrast to the inverted pendulum's arc-like movements of walking, running is a bouncing gait with sinusoidal vertical movements. A useful model of running simplifies the upper body to a point mass at the hip and each leg as a single linear spring. This spring-mass system lands with high kinetic and gravitational potential energy that is then converted into elastic energy stored in the spring. The spring rebounds during the second half of the stance phase and the elastic energy is converted back into kinetic and gravitational potential energy. The springs in a running person have been shown to be primarily the tendons of the leg extensor muscles. In reality, the musculoskeletal springs (tendons) are not perfect springs, they are about 93% elastic and so the muscles must shorten and perform work to replace the lost energy. More importantly, throughout the stance phase of running, the muscles must be active and maintain tension to allow the function of the spring-like tendons. Thus, running involves some muscle activity that performs work on the center of mass but a much greater muscle activity is needed for force generation to operate the springs and support body weight. The classic paper on both pendulum and spring-like gaits is Cavana, Heglund and Taylor, 1977 (1).

Although muscles from small animals and large exert equal stress, and perform mechanical work with equal efficiency, they differ in an important energetic measure, the economy of force production. If one stimulates a slow and a fast muscle fiber with equal length and cross sectional areas they will both produce the same isometric force. However, the faster fiber will consume metabolic energy at a much greater rate. This is because the cross bridges which produce force have a faster intrinsic rate of cycling and one molecule of ATP is consumed per cross bridge cycle. A longer fiber consumes more energy per unit force generated because a greater number of cross bridges cycle, i.e. attach, ratchet producing force, detach and then re-attach. Thus, the rate of energy consumption by a muscle generating isometric tension is proportional to its intrinsic crossbridge cycling rate and the volume of muscle (14). If muscles primarily act to generate force to maintain a rigid support leg in walking or to operate the tendon spring system, then the metabolic cost of locomotion would be proportional to the volume of active muscle the average intrinsic crossbridge cycling rate of that muscle volume.

We formalized and tested this "cost of generating force" hypothesis across both speed and animal size (9). We hypothesized that the metabolic rate of a running animal is proportional to the average force exerted on the ground (= body weight) and the rate of generating that force as indicated by the time an individual foot is in contact with the ground (t_c). The cost of generating force hypothesis is consistent with *in vitro* muscle experiments where the economy of force generation is directly related to the intrinsic rate of shortening of a muscle fiber. In other words, generating a unit force is more expensive in smaller animals because they use muscle fibers with faster intrinsic velocities. We proposed a simple equation,

$\dot{E}_{metab}/W_b = c \bullet 1/t_c$ where $\dot{E}_{metab}$ is the metabolic rate, W_b is body weight, c is a "cost coefficient" and t_c is the time of an individual foot contact per stride. If the hypothesis is correct, (i.e. if the metabolic rate is proportional to the average magnitude and rate of force development), then the cost coefficient would be a constant. We tested the hypothesis across a large size range of animals from small rodents to horses across their entire aerobic trotting and galloping speed range. The cost coefficient was nearly the same across this wide zoological comparison and within a given species the cost coefficient increased by less than 20% across a more than 300% range of running speeds (see Fig. 3).

Fig. 3



Optimal Tuning of Locomotion

Now that I have established the likely mechanisms that set the metabolic cost of animal locomotion, I want to consider the evidence that animals optimize their locomotion to minimize metabolic cost. The most fundamental descriptor of locomotion is "gait", that is the pattern of leg movements and footfall order. Bipeds, like humans, use primarily two gaits: the walk and the run. We naturally walk at slow speeds and run at fast speeds, but what about at intermediate speeds where we can either walk or run? At least four studies have demonstrated that in normal conditions, humans naturally choose the gait that consumes the least metabolic energy (8, 10, 12, 13) (see Figure 4). Quadrupedal animals use three basic gaits, the walk, the trot and the gallop. A classic study by Hoyt and Taylor (7) suggested that any each speed animals choose the energetically optimal gait. However, a more recent study by Farley and Taylor (3) indicated that the trigger for gait transitions are not energetic but in response to a critical level of musculoskeletal force. However, at most speeds animals do choose the gait that minimizes their metabolic cost.

Fig.4

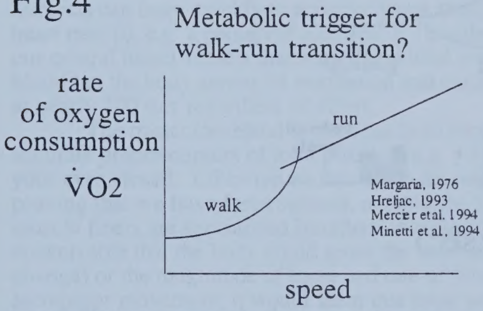
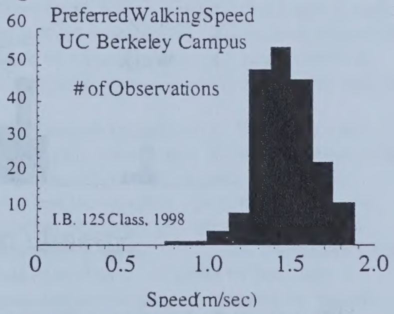


Fig. 5



A second fundamental way to characterize locomotion is velocity. Animals including humans display strong preferences for narrow speed ranges, although they could choose to locomote anywhere along the speed continuum. Humans prefer to walk at a speed of about 1.4 m/sec . Each year I have my undergraduate class collect preferred speed data from people traversing the Berkeley campus. These data are shown in Fig. 5. As you can see, the vast majority of individuals prefer to walk in a narrow band of speeds. What about the energetics of human walking? If we calculate the metabolic cost of transport, that is the Joules of energy per kilogram body mass per meter of forward distance traveled, we find a minimum at nearly 1.4 m/sec (10)(Fig. 6). Horses have also been shown to have preferred speeds within each gait and that the metabolic cost of transport is minimized near those preferred speeds (7) (Fig. 7).

Fig. 6

Cost of Transport - Humans

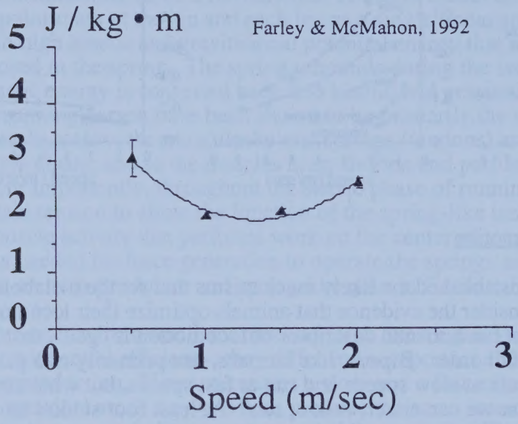
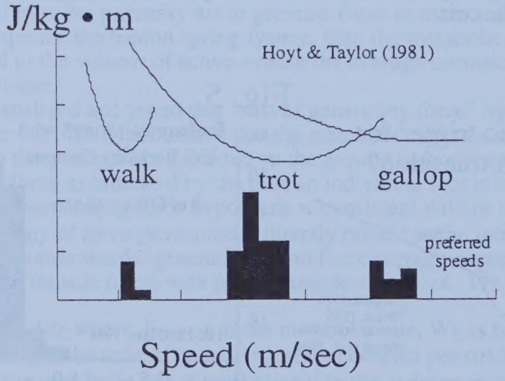


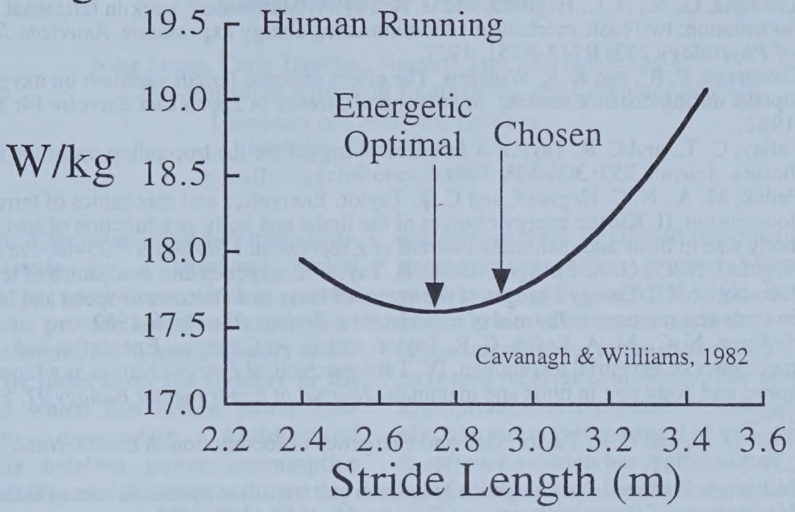
Fig. 7

Cost of Transport - horses



What about the kinematics or movement pattern at a particular locomotor speed? In one of the clearest studies, Cavanagh and Williams (2) showed that at a given running speed, people choose a combination of stride frequency and stride length that are nearly energetically optimal (Fig. 8). To determine the optimal stride frequency, they had their subjects run on a treadmill and match their stride frequency to a metronome set to a range of frequencies. They simultaneously measured their rate of oxygen consumption. Thus, there seems to be ample evidence that humans and other animals choose the speed, gait and kinematics that minimize their metabolic cost of locomotion.

Fig. 8



Biological Sensors and Software

In order for a locomotor movement to be controlled so as to minimize metabolic energy cost, it would seem that some biological sensors must exist. There are no fast acting direct sensors of metabolic energy consumption rate. However, there do seem to be many indirect sensors and a reasonably good method of integrating and processing these data at the higher cognitive levels. Psychophysiologists have documented that we are able to consciously rate our overall physical exertion. The Borg scale for perceived exertion is a scale from 6 to 20. Subjects can learn quickly to accurately rate their effort level with a number that corresponds to heart rate/10, e.g. a perceived exertion of 10 reflects a heart rate of 100. There are sensors in our central blood vessels that sense the partial pressure of oxygen dissolved in our arterial blood but the body servos its ventilation and cardiac output to maintain arterial oxygen tension at nearly 100 torr regardless of effort.

The musculoskeletal system has both force and length transducers. We have very accurate proprioceptors of joint position, e.g. you can touch your finger tip to your nose with your eyes closed. Likewise we can sense the weight of an object as we pick it up, thus proving that we have force sensors, notably the Golgi tendon organs. Intermingled in our muscle fibers are specialized intrafusal muscle fibers that sense length change. Thus, it is conceivable that the body could sense the work performed (product of muscle force and length change) or the magnitude of force and rate of force development. In order to fine tune a locomotor movement, it would seem that local sensors and some overall integrating neural software would need to exist. Research in this area is in its infancy.

Summary

Animals produce locomotion using muscles as actuators and important mechanical energy conserving mechanisms: pendulums in walking and springs in running. As a result, it is the economical production of muscular force that sets most of the rate of metabolic energy consumption. Phenomenologically, there is ample evidence that animals locomote at the speeds, gaits and kinematics that minimize metabolic cost. We have a poor understanding of how biological sensors provide feedback that could result in such optimization.

References

1. Cavagna, G. A., N. C. Heglund, and C. R. Taylor. Mechanical work in terrestrial locomotion: two basic mechanisms for minimizing energy expenditure. *American Journal of Physiology* 233: R243-R261, 1977.
2. Cavanagh, P. R:, and K. R. Williams. The effect of stride length variation on oxygen uptake during distance running. *Medicine and Science in Sports and Exercise* 14: 30-5, 1982.
3. Farley, C. T., and C. R. Taylor. A mechanical trigger for the trot-gallop transition in horses. *Science* 253: 306-308, 1991.
4. Fedak, M. A., N. C. Heglund, and C. R. Taylor. Energetics and mechanics of terrestrial locomotion. II. Kinetic energy changes of the limbs and body as a function of speed and body size in birds and mammals. *Journal of Experimental Biology* 97: 23-40, 1982.
5. Heglund, N. C., G. A. Cavagna, and C. R. Taylor. Energetics and mechanics of terrestrial locomotion. III. Energy changes of the centre of mass as a function of speed and body size in birds and mammals. *Journal of Experimental Biology* 97: 41-56, 1982.
6. Heglund, N. C., M. A. Fedak, C. R. Taylor, and G. A. Cavagna. Energetics and mechanics of terrestrial locomotion. IV. Total mechanical energy changes as a function of speed and body size in birds and mammals. *Journal of Experimental Biology* 97: 57-66, 1982.
7. Hoyt, D. F., and C. R. Taylor. Gait and energetics of locomotion in horses. *Nature* 292: 239-240, 1981.
8. Hreljac, A. Preferred and energetically optimal gait transition speeds in human locomotion. *Medicine and Science in Sports and Exercise* 25: 1158-1162, 1993.
9. Kram, R., and C. R. Taylor. Energetics of running: a new perspective. *Nature* 346: 265-267, 1990.
10. Margaria, R. *Biomechanics and Energetics of Muscular Exercise*. Oxford: Clarendon Press, 1976.
11. McMahon, T. A. *Muscles, Reflexes, and Locomotion*. Princeton, NJ: Princeton University Press, 1984.
12. Mercier, J., D. Le Gallais, M. Durand, C. Goudal, J. P. Micallef, and C. Prefaut. Energy expenditure and cardiorespiratory responses at the transition between walking and running. *European Journal of Applied Physiology* 69: 525-529, 1994.
13. Minetti, A. E., L. P. Ardigo, and F. Saibene. The transition between walking and running in humans: metabolic and mechanical aspects at different gradients. *Acta Physiologica Scandanavica* 150: 315-323, 1994.
14. Rall, J. A. Energetic aspects of skeletal muscle contraction: implications of fiber types. *Exercise and Sport Science Reviews* 13: 33-74, 1985.
15. Taylor, C. R., N. C. Heglund, and G. M. Maloiy. Energetics and mechanics of terrestrial locomotion. I. Metabolic energy consumption as a function of speed and body size in birds and mammals. *Journal of Experimental Biology* 97: 1-21, 1982.
16. Taylor, C. R., N. C. Heglund, T. A. McMahon, and T. R. Looney. Energetic cost of generating muscular force during running: a comparison of small and large animals. *Journal of Experimental Biology* 86: 9-18, 1980.

Architectural Implementation Issues in a Wideband Receiver Using Multiuser Detection

Ning Zhang, Craig Teuscher, Hungchi Lee, Bob Brodersen
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720, USA
e-mail: ningzh@eecs.berkeley.edu

Abstract: This paper considers implementation issues for a baseband digital receiver in an indoor wireless system which can provide multimedia access to portable terminals in high capacity situations. The focus is on the receiver in the terminal which has severe power consumption constraints. Architectural trade-offs between power consumption and flexibility are discussed with particular focus on the circuitry required to support adaptive multiuser detection.

1. Introduction

The InfoPad project at UC Berkeley provides an experimental example of a wireless system requiring both high performance and low-power operation [1,2]. A recent focus has been to provide wireless multimedia access to as many as 15 portable terminals at data rates up to 3.3 Mbit/sec per terminal within each cell of an indoor, picocellular environment.

A first-generation design for the InfoPad application has been previously reported in [3]. That work describes a two chip, integrated CMOS receiver implementation which achieves target speed and power dissipation levels. However, receiver performance is severely degraded in the presence of multiple access interference (MAI). Losses are a direct result of the simple matched filter correlator used for signal detection [4].

Multiuser detection provides a potential solution to this problem. Multiuser detectors achieve significant performance gains over conventional receivers when performance is limited by MAI [5]. Benefits are achieved by exploiting the structure of the interference, at the expense of increased receiver complexity. The most appropriate implementation of this complex algorithm is the subject of this paper. A software solution has the benefit of being closer to the simulations and allowing more flexibility, whereas a dedicated hardware solution is more energy efficient.

A complete system implementation of the receiver is being pursued which includes both analog and digital components. This design will be quickly reviewed with an in depth discussion of the digital baseband algorithm and possible architectures.

2. Algorithm

System Specification

Direct-sequence code division multiple access (DS-CDMA) is used for multiplexing system users. Individual spreading sequences are constructed using cyclic shifts of a single prime MLSR sequence of length 15, with one sequence reserved for pilot channel signaling. Quadrature phase shift keying (QPSK) is used for data modulation with a symbol rate of 1.67 MHz (25 MHz chip rate). At the transmitter, a 25% excess bandwidth raised-cosine filter restricts spectral occupancy to approximately 32 MHz at a carrier frequency of 2

GHz, near the likely spectral location for a future wideband CDMA standard. Independent transmit power control levels (adjustable from 0-10 dB) are available on each data channel, and an aggregate transmit power of 0 dBm is assumed. The picocellular coverage area for each base station is approximately 10 meters in diameter.

A key feature of this design is the use of a dedicated pilot channel to facilitate timing recovery, gain control, channel estimation, and multiuser detection. Normally, pilot channels are not used because they are wasteful of bandwidth; decision-directed techniques are preferred instead. For the InfoPad system, however, bandwidth allocated to the pilot channel is amortized over the total number of active users in the cell. Since data is aggregated and synchronized prior to transmission, all mobile receivers in a cell can share a single pilot tone, thus reducing the effective pilot overhead. Because receiver operation depends on the presence of the pilot signal, the power level for this signal is assumed to be at least equal to the highest power level of any other active user.

A "mostly digital" radio architecture serves as the basic model for this design [6]. Because digital circuits are particularly amenable to low-power implementation, a concerted effort is made to "go digital" at the earliest possible point in the receive chain. Analog hardware simplifications are favored, even at the expense of increased digital baseband complexity. Two custom CMOS integrated circuits are proposed for this application: one to implement all RF and analog baseband receiver functions, and one to implement the digital baseband processing. The design of the whole wideband spread spectrum radio is described in [7]. This paper focuses on the implementation of the digital baseband section.

Digital Baseband Section

The digital baseband section of the receiver is required to perform the following functions: timing recovery, channel estimation, RAKE matched filtering, adaptive multiuser detection, and data recovery. Figure 1 shows a block diagram of the proposed design.

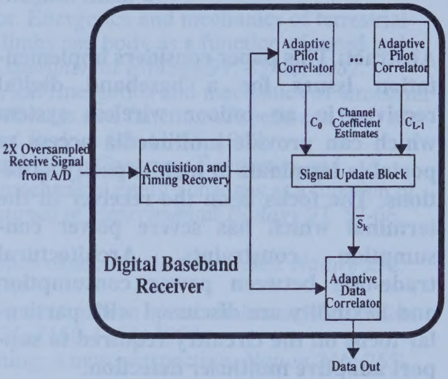

Adaptive Pilot Correlator
...
Adaptive Pilot Correlator
C_0
Channel Coefficient Estimates
C_{L-1}
2X Oversampled Receive Signal from A/D
Acquisition and Timing Recovery
Signal Update Block
$\overline{S}_k$
Digital Baseband Receiver
Adaptive Data Correlator
Data Out

Figure 1. Block diagram of the digital baseband receiver.

Timing recovery in the receiver is accomplished in two stages: initial coarse signal acquisition, followed by continuous pilot channel tracking. The acquisition algorithm relies on an exhaustive serial search over all code phase offsets of the pilot sequence. The optimum phase for the local sequence is selected with a resolution of one half chip time. Following acquisition, track mode circuitry implements a digital delay locked loop (DLL) to attain and maintain synchronization with the pilot channel to within one quarter chip time. Tracking circuitry also monitors a loss-of-lock condition to initiate prompt signal reacquisition, when necessary.

The presence of a pilot signal at the receiver also aids in the formation of reliable channel estimates. Coefficient estimates are necessary in order to implement the RAKE matched filter described below.

Channel estimates are formed using a bank of adaptive pilot correlators, one for each coefficient c_k. Each correlator tracks a single multipath component of the pilot signal. Like the data correlator described below, the pilot correlator is subject to severe MAI and therefore requires adaptive multiuser detection also. Correlator adaptation follows the method proposed in [8], a straightforward extension of BAMUD to multipath settings. And adaptation is accomplished using DLMS algorithm [9] with D=2. Rapid correlator convergence is facilitated by the absence of data modulation on the pilot signal. "Blind" adaptation is not required since a training sequence is constantly available.

The signal update block performs channel coefficient averaging, signal amplitude estimation, and RAKE matched filter construction. Channel coefficient averaging improves the accuracy of channel estimates as long as the channel remains approximately static during the estimation interval. Pilot channel amplitude estimates are also formed by averaging. For this implementation, it is assumed that the receiver has knowledge of the desired signal amplitude relative to the pilot channel amplitude so that the receiver only needs to form the pilot signal amplitude estimate. The final task of the signal update block is to generate an estimate of the effective signature sequence $\bar{s_k}(t)$ using the most recent channel coefficient estimates. To the extent that these estimates are accurate, the resulting signal constitutes the RAKE matched filter to the desired user's signal, providing Lth-order receiver diversity to combat signal fading.

The adaptive data correlator performs data recovery which also implements adaptive multiuser detection using the DLMS algorithm. Three different modes of adaptation can be implemented: training sequence based, decision-directed, and blind. External control signals permit the receiver to dynamically switch between these three modes of operation. As noted previously, the anchored component of the adaptive correlator is generated by the signal update block. The update rate for this block is user-programmable.

3. Architecture

The significant performance gains of the proposed receiver over conventional matched filter receiver are achieved by increased digital baseband processing complexity. To implement this algorithm in an area efficient, low power integrated circuit requires a careful consideration of the architectural choice.

The major trade-off involved in architecture design is flexibility vs. efficiency. Two extremes which will be compared here are the direct mapping approach and a software programmed processor. In a direct mapped implementation, each algorithmic function has a corresponding dedicated hardware unit thus providing a highly parallel implementation. Energy efficiency is achieved by voltage reduction (which is possible because of the high degree of parallelism) and by the high level of optimization of the hardware for a particular algorithm.

Programmable digital signal processors (DSP) are at the other extreme, where a variety of algorithms can be implemented on the same processor at the expense of larger energy consumption. There are other alternative architectures in between, such as hardware reconfigurable implementations and programmable processors enhanced with dedicated accelerators. A typical reconfigurable approach makes use of FPGA's (field programmable gate arrays), where the algorithm can be direct mapped onto configurable logic blocks. Such implementations have more flexibility than a direct mapped approach, but at the expense of higher power consumption and area. The use of dedicated accelerators allows more efficiency for those functions which are optimized, but the flexibility is reduced as the approach becomes more dedicated.

In the following sections, an architectural comparison is made of two architectural extremes to bound the cost of flexibility in terms of area and power. To this end a fully dedicated direct mapped implementation and the software specified DSP approach are considered based on the algorithm described in section two.

Computational Complexity

Since adaptive pilot correlators are the major blocks of the digital baseband receiver, and since it has almost identical computation as the adaptive data correlator, the comparison in the following sections are based on one adaptive pilot correlator.

The received complex data comes in at symbol rate of 1.67 MHz. For each symbol, the dominant computation inside a single adaptive correlator are two multiply accumulate operations of length 15, one with real coefficients and the other with complex coefficients and thirty multiplications, fifteen with one input real and one input complex, and the other fifteen with both inputs complex. In addition, the adaptive pilot correlator needs to perform thirty-three complex additions/subtractions and one complex shift. Count all the operations together, 300 million equivalent real multiplications and 357 million real additions/subtractions per second are required for this single correlator.

4. DSP Implementation

The Texas Instruments TMS320C54x is a family of 16-bit fixed-point DSP's which has on-chip memories and an enhanced Harvard architecture. TMS320LC54x is designed specifically for low-power operation. The TMS320C549 has the lowest energy consumption for all the eleven BDT DSP Benchmark [10] when compared with nineteen commonly used DSP's, with an energy consumption of less than 50% of its closest competitor.

C code describing the function of one adaptive pilot correlator was compiled and linked using the complier and linker provided at TI's "Digital Signal Processing DSP On-line Lab". Performance and power calculations are based on benchmarking compiled assembly codes and adding the contribution of all instructions using TMS320LC54x instruction level power consumption data measured and published by Texas Instruments [11]. Initialization section of the program was excluded in the performance and power calculation. (A 3V supply voltage and 40 MHz clock rate were used in the calculations.) The results show that the total power consumption is 4 W and 56 processors would be required to achieve real time operation.

TI's newly developed low power DSP for wireless communications, which operates from 1 V supply and is fabricated in a 0.35 um dual-Vt technology with 0.25 um minimum gate lengths, achieves 63 MHz operation with 0.21mW/MHz [12]. Since its architecture is based on TMS320C54x, and assuming it takes the same number of cycles to execute the program to implement an adaptive pilot correlator, then the power it consumes is 460 mW and 35 parallel processors are required to achieve the required data throughput.

It was found that arithmetic instructions only take about 30% of the total execution time and power, while memory accesses and control take the majority of both. This indicates that considerable advantages might be achieved in a direct mapped design, which has duplicated functional units with dedicated data sources and vastly reduced control.

5. Direct Mapped Implementation

Though it is well known that dedicated hardware maximizes energy efficiency, it is the amount of improvement which is being investigated here. In addition this implementation serves as an indicator of the complexity achievable with current semiconductor technology when there is significant power and cost constraints. The integrated circuit fabrication is planned in a standard 0.25 um digital CMOS process. Figure 2 shows a block diagram of the direct-mapped implementation of an adaptive pilot correlator.

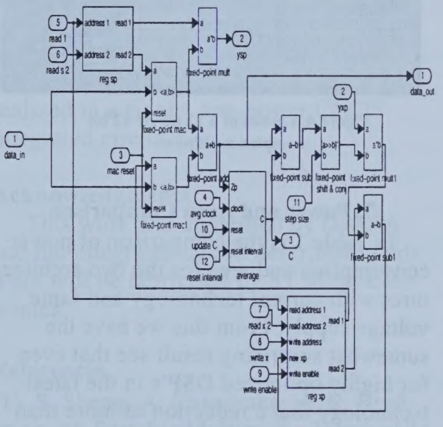

5
read 1
6
read s 2
address 1
read 1
address 2
read 2
reg sp
a
b
a*b
2
ysp
fixed-point mult
a
b <a,b>
reset
fixed-point mac
1
data_in
3
mac reset
a
b <a,b>
reset
fixed-point mac1
a
b
a+b
fixed-point add
Zp
clock
reset
reset interval
C
average
4
avg clock
10
update C
12
reset interval
1
data_out
2
yxp
a
b
a-b
fixed-point sub
a
b
(a>>b)'
fixed-point shift & conj
11
step size
3
C
a
b
a*b
fixed-point mult1
a
b
a-b
fixed-point sub1
7
read x 2
8
write x
9
write enable
read address 1
read address 2
write address
new xp
write enable
read 1
read 2
reg xp

Figure 2. Block Diagram of Adaptive Pilot Correlator.

Finite Word Length Effects

One important design choice in a direct map implementation is the bit width of the arithmetic blocks. Power and area considerations motivate the use of minimum word widths, while algorithmic accuracy and performance favor longer words. The LMS algorithm is relatively robust in the presence of finite precision effects, and thus significant savings can be achieved.

Simulations on receiver performance with finite word length effects have been performed. The simulations assume that a 10 bit A/D converter generates the input data stream, thirteen users are active, and the desired user transmits at the smallest possible (non-zero) power level while all other users transmit at the maximum. The simulation results show that truncation at the output of multiplier and multiply-accumulator (MAC) degrades correlator performance more than rounding due to the DC offset introduced in the LSB that accumulates over time. In this simulation, 16 bit arithmetic was found to be essentially indistinguishable from floating point results, with a small performance reduction with 12 bits.

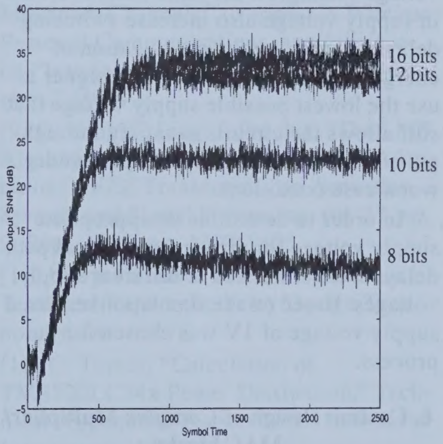


Figure 3. Finite Word Length Effect.

For the implementation, truncation is preferable since it simplifies hardware design and reduces power consumption. Further investigations on the performance degradation resulted from truncation shows that the main problem lies in the error signal generation, where the absolute value of the signal can be extremely small after convergence. This degradation can be eliminated by implementing a single rounding operation after the error signal multiplication. Simulation results for this scenario are illustrated in Figure 3, which are virtually indistinguishable from the results when rounding is performed in all arithmetic blocks. Using error signal rounding only, the performance penalty for 12 bit arithmetic becomes noticeable only when the output SIR for the receiver exceeds the high value of 30 dB. Thus 12 bit fixed point arithmetic was used.

Supply Voltage

One of the principal techniques for reducing power consumption in digital circuits is to lower the supply voltage. Because dynamic power consumption is proportional to the square of the supply voltage, quadratic improvements in power dissipation can be realized through supply voltage scaling. At the same time, decrease in supply voltage also increase switching delays. Consequently, optimization of energy efficiency requires the designer to use the lowest possible supply voltage that still allows the circuits to satisfy throughput constraints on the critical path under worst case conditions.

In order to determine an appropriate supply voltage for this design, critical path delays were simulated at different supply voltages. Based on the simulation results, a supply voltage of 1V was chosen for the process.

6. Custom Design of Complex Multiplier/ MAC blocks

The overall size and power consumption of the direct-mapped implementation are dominated by the complex multipliers and MAC's required for algorithm implementation. Consequently, fully custom, low-power, 12 bit by 12 bit complex multiplier and MAC blocks have been designed specifically for this application in order to achieve a low-power and small area implementation.

The final multiplier design is a parameterized parallel structure with two pipeline stages. It has area of 0.18mm^2 and power consumption of 0.62mW with 1V voltage supply and 25Mhz clock. Figure 4 depicts the layout of the complex multiplier.

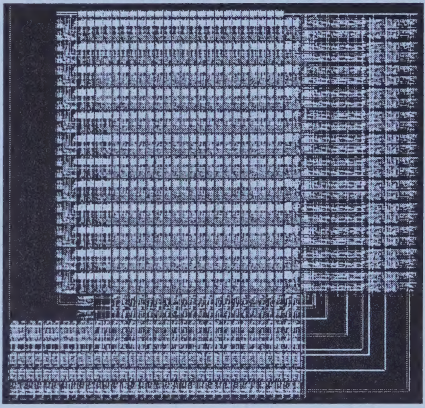


Figure 4. Layout of a 12 bit by 12 bit complex multiplier.

7. Power and Area Comparison

In Table 1 is the comparison of power consumption and area for the two architectures with current technology and same voltage supply. From this we have the somewhat surprising result see that even for highly optimized DSP's in the latest technology that a reduction of more than two orders of magnitude in power (150 times) and area (> 700 times) is possible with the direct mapped approach. The reasons for these high ratio's are now under investigation.

Table 1: Power and Area Comparison

	DSP Implementation	Direct Mapped Implementation
Adaptive Pilot Correlator	460 mW 1089 mm^2	3 mW 1.3 mm^2
Digital Baseband Receiver	1500 mW 3600 mm^2	10 mW 5 mm^2

8. Conclusions

This paper describes the implementation of a wideband baseband digital receiver using adaptive multiuser detection in an indoor wireless environment. Two possible architectures have been considered and compared. In order to achieve the targeted power consumption level, a direct mapped architecture is more suitable with current technology. In spite of the additional complexity required by implementing multiuser detection to suppress MAI, the entire digital baseband section can be realized in a single, low-power CMOS integrated circuit.

Acknowledgments

This work was supported by DARPA and the integrated circuit described in this work will be fabricated by ST MicroElectronics.

References

[1] S. Sheng, A. Chandrakasan, R. Brodersen, "A Portable Multimedia Terminal," *IEEE Comm. Mag.*, Dec. 1992.

[2] S. Sheng, Low-Power CMOS Wireless Communications, Boston: Kluwer, 1998.

[3] S. Sheng, et al, "A Low Power CMOS Chipset for Spread Spectrum Communications," IEEE Intern. Solid-State Circuits Conf., Feb. 1996.

[4] J. Linnartz, L. Yun, M. Couture. "Multi-user and Self Interference Effects in a QPSK DS-CDMA Downlink with a Ricean Dispersive Channel," Wireless Personal Communications, vol. 1, 1995, pp. 95-102.

[5] R. Lupas, S. Verdu, "Linear Multiuser Detectors for Synchronous Code-Division Multiple Access Channels," IEEE Trans. Inform. Theory, vol. IT-35, Jan. 1989.

[6] C. Teuscher, et al, "Design and Implementation Issues for a Wideband, Indoor, DS-CDMA System Providing Multimedia Access," Proceedings of the Thirty-Fourth Annual Allerton Conference, Oct. 1996, pp. 623-632.

[7] C. Teuscher, D. Yee, N. Zhang, R. Brodersen, "Design of a Wideband Spread Spectrum Radio Using Adaptive Multiuser Detection," ISCAS, June 1998.

[8] H. Huang, S. Verdu, "Linear Differentially Coherent Multiuser Detection for Multipath Channels," to appear in Wireless Personal Communications, special issue on "Interference in Moble Wireless Systems," 1998.

[9] G. Long, F. Ling, J. Proakis, "The LMS Algorithm with Delayed Coefficient Adaptation," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 9, Sept. 1989, pp. 1397-1405.

[10] "Buyer's Guide to DSP Processors," Technical Report, Berkeley Design Technology, Inc.

[11] C. Turner, "Calculation of TMS320LC54x Power Dissipation," Technical Application Report SPRA164, Texas Instruments, 1997.

[12] W. Lee et al., "A 1-V Programmable DSP for Wireless Communications," IEEE Journal of Solid-state Circuits, vol. 32, no. 11, Nov. 1997, pp. 1766-1777.

Iterative Techniques for DS/CDMA Multipath Channel Estimation

Emre Ertin — ertine@er4.eng.ohio-state.edu
Urbashi Mitra — ubli@er4.eng.ohio-state.edu
Siwaruk Siwamogsatham — siwamogs@er4.eng.ohio-state.edu
Department of Electrical Engineering,
The Ohio State University, Columbus OH 43210

Abstract

In this paper, the problem of estimating the channel parameters of a new user in a multi-user Direct-Sequence/Code-Division Multiple-Access communication system is addressed. It is assumed that the new user transmits training data over a slowly fading multipath channel. The proposed algorithm is based on maximum likelihood estimation of the channel parameters. First, an asymptotic expression for the likelihood function of the channel parameters is derived and a re-parametrization of this likelihood function is proposed. In this re-parametrization, the channel parameters are combined into a discrete time channel filter of symbol period length. Then, Expectation Maximization and Alternating Projection based techniques are considered to extract channel parameters from the estimated discrete channel filter, to maximize the derived asymptotic likelihood function. The performance of the proposed algorithms is evaluated through simulation studies as well as through comparison with the derived Cramér-Rao bounds.

I. Introduction

Direct-Sequence Code-Division Multiple-Access (DS-CDMA) is emerging as a possible multiple access scheme for future digital wireless communication systems. The inherent low power, potential for high capacity, anti-jamming and anti-multipath characteristics of DS-CDMA systems motivate its consideration. While a large number of receivers for multi-user DS-CDMA have been proposed, a significant portion of these proposals rely on knowledge of communication parameters. Specifically, an accurate channel description is often necessary. The communications channel of each user in a multipath channel can be modeled by a tapped-delay line [1] and as such, the necessary parameter knowledge is the set of time delays and coefficients characterizing each path of the channel.

In this paper, maximum-likelihood (ML) based schemes are considered for channel estimation of a multi-user DS-CDMA system in a multipath environment. It is assumed that the unknown path delays are continuously valued. Initially, the classical ML estimation problem is posed where the multiple-access interference (MAI) is modeled as colored Gaussian noise as in [2, 3]. The asymptotic statistics for the distribution of the MAI interference are derived. In particular, it is shown that for sufficiently long observation intervals, the interference subspace has essentially dimension K, where K is the number of active users. For the multipath scenario, it is observed that such a scheme remains prohibitively complex. In order to reduce computational complexity, two iterative techniques techniques based on the expectation-maximization (EM) [4] and the alternating projection (AP) [5] algorithms are presented.

It should be explicitly noted that due to the investigation of the maximum-likelihood criterion, the work presented herein has similarities to prior work [6, 3, 7, 8]. The focus of the current work is on channel estimation for a new single-user transmitting in a multipath, multi-access environment where no prior synchronization is assumed. The contribution of this work is in the derivation and use of asymptotic statistics which have a structure that can be exploited to simplify the estimation algorithm and the development of an alternative parameterization of the signal of interest, which facilitates the use of computationally efficient iterative techniques for the complex ML optimization problem. These developments directly impact estimator form and improve performance. Furthermore, the asymptotic statistics enable interpretation of the operation of the channel estimators.

This paper is organized as follows. In Section II, the system model is introduced. In Section III, we provide the derivation of the asymptotic likelihood function. Estimation of the channel parameters via exact ML and suboptimal approximations of the ML method is presented in Section IV. Section V-VI present simulation results and conclusions.

II. System Model

We consider an asynchronous DS-CDMA system operating over a slow fading multipath channel. The baseband signal for user k is obtained by spreading a BPSK modulated data stream $b_k(i)$ onto a spreading waveform $s_k(t)$. The support of $s_k(t)$ is $[0, T)$, where T is the symbol period. The spreading waveform is generated by modulating a signature sequence $\{c_k(n)\}_{n=0}^{N-1}$ (m-sequence, Gold Code, etc. [9]) of length N with a train of rectangular pulses of duration $T_c = T/N$. The transmitted signal is formed by multiplying the baseband signal with a carrier of amplitude A_k and phase ϕ_k. The complex envelope representation of the transmitted signal is given by

$$x_k(t) = \sum_{i=-\infty}^{\infty} A_k e^{j(\omega_c t+\phi_k)} s_k(t-iT) b_k(i). \tag{1}$$

We assume that the channel for user k consists of L_k distinct, resolvable, propagation paths [10]. The impulse response $h_k(t)$ of the channel is given by

$$h_k(t) = \sum_{l=1}^{L_k} h_{k,l}\delta(t-\tau_{k,l}), \tag{2}$$

where $h_{k,l}$ is the complex channel fading coefficient and $\tau_{k,l}$ is the propagation delay associated with the l^{th} propagation path for k^{th} user's signal. It is assumed that $h_{k,l}$ and $\tau_{k,l}$ are unknown, but do not vary throughout the period of consideration. A maximum multipath spread of half the symbol period is assumed; i.e, $\max(\tau_{k,l}) - \min(\tau_{k,l}) < T/2$. Then, without loss of generality, we can assume that the support of $h_k(t)$ lies in the interval $[0, 3T/2)$.

At the receiver, the received signal is the superposition of propagated signals from all users and the background channel noise. The complex envelope representation of the received signal $r(t)$ can be written as

$$r(t) = \sum_{k=1}^{K} h_k(t) * x_k(t) + n(t) = \sum_{k=1}^{K}\sum_{l=1}^{L_k}\sum_{i=-\infty}^{\infty} \alpha_{k,l} e^{j\omega_c t} s_k(t-iT-\tau_{k,l}) b_k(i) + n(t),$$

where $*$ denotes the convolution operation, $\alpha_{k,l} = A_k |h_{k,l}| e^{j(\phi_k - \omega_c \tau_{k,l} + \angle h_{k,l})}$, and $n(t)$ is the additive white complex Gaussian noise with zero mean and a power spectral density $\mathcal{N}_0$. After the carrier is suppressed, a discrete-time observation is formed by sampling the output of a filter matched to the chip waveform at the chip rate. As the channel parameters are unknown, the chip matched filter is not synchronized with any of the

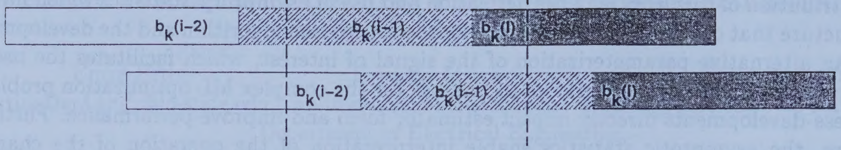


Figure 1: Signal vectors for a user transmitting through a multipath channel with two paths users' symbol intervals or chip transitions. The received vector $\underline{y}_i$ is formed by stacking N samples associated with the i^{th} bit interval: $\underline{y}_i = [\, y(iN)\; y(iN+1)\; \ldots\; y(iN+N-1)\,]^T$.

The contribution of the K active users to the i^{th} observation vector $\underline{y}_i$ can be written as

$$\underline{y}_i^{MAI} = \sum_{k=1}^{K} \sum_{l=1}^{L_k} \alpha_{k,l} \left(b_k(i-2)\underline{s}_k^l(\tau_{k,l}) + b_k(i-1)\underline{s}_k^m(\tau_{k,l}) + b_k(i)\underline{s}_k^r(\tau_{k,l}) \right) + \underline{n}_i, \tag{3}$$

where $\underline{s}_k^l(u)$, $\underline{s}_k^m(u)$, and $\underline{s}_k^r(u)$ are defined as follows. Let $u \in \{0, 1, \ldots, N + \lfloor N/2 \rfloor\}$ and $\gamma \in [0, 1)$ respectively be the integer part and the fractional part of the normalized delay τ/T_c, *i.e.*, $\tau = (u + \gamma)T_c$, where $\lfloor N/2 \rfloor$ denotes the largest integer less than or equal to $N/2$. When the delay τ is perfectly aligned with the sampling time ($\tau = uT_c$), $\underline{s}_k^l(u)$, $\underline{s}_k^m(u)$, and $\underline{s}_k^r(u)$ are given by

$$\underline{s}_k^l(\tau) = \begin{cases} \underline{0} & u \leq N \\ [\, c_k(2N-u) \ldots c_k(N-1)\; 0 \ldots 0\,]^T & \text{otherwise} \end{cases}$$

$$\underline{s}_k^m(\tau) = \begin{cases} \underline{0} & u = 0 \\ [c_k(N-u) \ldots c_k(N-1)\; 0 \ldots 0]^T & 0 < u \leq N \\ [0 \ldots 0\; c_k(0) \ldots c_k(2N-1-u)]^T & \text{otherwise} \end{cases}$$

$$\underline{s}_k^r(\tau) = \begin{cases} [0 \ldots 0\; c_k(0) \ldots c_k(N-1-u)]^T & 0 \leq u < N \\ \underline{0} & \text{otherwise.} \end{cases}$$

Since $\underline{y}_i$ is obtained from an integrate-and-dump circuit, the spreading vectors for a general τ will be a convex combination of two adjacent chips; *i.e.*, $\underline{s}_k^x(\tau) = (1-\gamma)\underline{s}_k^x(u) + \gamma \underline{s}_k^x(u+1)$, where $x \in \{l, m, r\}$. For the single-path case, an arbitrary block of observations with length equal to the symbol interval (length N) will contain at least the end of the previous symbol and the beginning of the current symbol of each user. Thus, this block can be viewed as a linear combination of 2K signal components ($s_k^r(\tau_k)$'s and $s_k^m(\tau_k)$'s) and noise. For the multipath case, the signal of each user which arrives from each propagation path may also superimpose upon each other. Thus, with our assumption of a maximum multipath spread, a block of observations can contain at most three adjacent bits of each user. Figure 1 illustrates an example where the channel of user i with two multipath components results in three signal vectors in one observation period. We can write the signal model in (3) in matrix vector form: $\underline{y}_i^{MAI} = S_{MAI} A_{MAI} \underline{b}_{MAI}(i) + \underline{n}_i$, where S_{MAI}, A_{MAI} and $\underline{b}_{MAI}(i)$ are defined by:

$$\begin{aligned} \underline{\alpha}_k &= [\alpha_{k,1} \ldots \alpha_{k,L}\,]^T,\; A_{MAI} = [\operatorname{diag}([\underline{\alpha}_1, \underline{\alpha}_1, \underline{\alpha}_1, \underline{\alpha}_2, \cdots, \underline{\alpha}_K])]_{(3KL \times 3K)}, \\ S_k &= [\; \underline{s}_k^l(\tau_{k,1}) \;\cdots\; \underline{s}_k^l(\tau_{k,L})\; \underline{s}_k^m(\tau_{k,1}) \;\cdots\; \underline{s}_k^m(\tau_{k,L})\; \underline{s}_k^r(\tau_{k,1}) \;\cdots\; \underline{s}_k^r(\tau_{k,L})\;], \\ S_{MAI} &= [S_1\; S_2\; \ldots\; S_M]_{(N \times 3KL)}, \\ \underline{b}_{MAI}(i) &= [b_1(i-2) b_1(i-1) b_1(i) \ldots b_M(i-2) b_M(i-1) b_M(i)]^T. \end{aligned}$$

Now consider the signal component of the new user $k = 0$, transmitting with spreading code $c_0(k)$. During the acquisition period, the new user will transmit a fixed-symbol training sequence of length J for initial acquisition. The remaining active users continue transmitting zero-mean random signals which will be treated as multiple access interference.

Since the data bits and the channel parameters of the new user are constant during this period, the propagation delay τ only causes a circular shift to the spreading vector c_0. For $u \in [0, \ldots, N-1]$ we define the circularly shifted spreading vector: $\underline{s}_k(u) = [\ c_k(N-u) \ldots c_k(0) \ldots c_k(N-u-1)\ ]^T$, and a signature matrix for user (k) can be obtained by stacking the circularly shifted versions of the spreading vector column wise: $C_k = [\underline{s}_k(0)\underline{s}_k(1) \ldots \underline{s}_k(N-1)]$. Then, the effective spreading code for a propagation delay τ is given by $C_0\delta(\tau)$, where $\delta(\tau)$ is defined for $\tau/T_c = u + \gamma$ as $\delta(\tau) = (1-\gamma)\delta^u + \gamma\delta^{u+1}$.

The model for $\underline{y}_i$ can be obtained by adding up the contribution of the active users with the constant signals from the new user. That is, $\underline{y}_i = C_0 D(\underline{\tau}_0)\underline{\alpha}_0 + S_{MAI}A_{MAI}\underline{b}_i + \underline{n}_i$, where the $N \times L$ matrix $D(\underline{\tau}_0)$ is given by $D(\underline{\tau}_0) = [\delta(\tau_{0,1}), \ldots, \delta(\tau_{0,L})]$. We define the $D(\underline{\tau}_0)\underline{\alpha}_0$ as the effective discrete channel filter $\underline{\tilde{h}}_0(\underline{\tau}_0, \underline{\alpha}_0)$ for user $k = 0$ and obtain an alternative representation for the received signal as:

$$\underline{y}_i = C_0\underline{\tilde{h}}_0(\underline{\tau}_0, \underline{\alpha}_0) + S_{MAI}A_{MAI}\underline{b}_{MAI}(i) + \underline{n}_i.$$

III. Asymptotic Likelihood Function

In this paper, we consider the problem of estimating the unknown channel parameters of a new active user based on observations made during an acquisition period. The desired parameters are the multipath propagation delays for L distinct paths and their associated complex valued attenuation coefficients. While it is possible to postulate a joint maximum-likelihood (ML) estimator to simultaneously estimate the $3KL$ (L paths for each of K users) real valued parameters, the resultant optimization is prohibitively complex. Our method will be to extend the techniques of [3] to the current multipath problem. In particular, we will reduce the joint ML estimation problem to a simpler single-user problem through the consideration of the sample mean, $\underline{m}_J$, of the observed random vectors $\{\underline{y}_i\}_{i=1}^{J}$. Recall that J is the length of the training sequence. We define $\underline{m}_J$ to be:

$$\underline{m}_J = \frac{1}{J}\sum_{i=1}^{J}\underline{y}_i = C_0\underline{\tilde{h}}_0(\underline{\tau}_0, \underline{\alpha}_0) + \frac{1}{J}\sum_{i=1}^{J}\left[S_{MAI}A_{MAI}b_{MAI}(i) + \underline{n}_i\right]. \tag{4}$$

Next, we show that the asymptotic distribution of the sample average $\underline{m}_J$ is Gaussian. The data bits of the interfering users are treated as random variables, whereas the channel parameters of all users are assumed to be deterministic variables.

While it is true that the $\underline{y}_i$ in (4) are identically distributed, these random vectors are not independent. However, they are m-dependent. That is, for $|m| \geq 3$, $\underline{y}_i$ and $\underline{y}_{i+m}$ are independent. This is due to the fact that each data bit of an active user can contribute to at most three possible adjacent observation intervals. We can invoke the Central Limit Theorem for m-dependent random variables [11]. Thus, the random variable $(\sqrt{J}\Sigma_J^{-1/2}(\underline{m}_J - \underline{\mu}))$ converges in distribution to $N(\underline{0}, I)$, where

$$\underline{\mu} = E\left[\underline{m}_J\right] = C_0\underline{\tilde{h}}_0(\underline{\tau}_0, \underline{\alpha}_0) \quad \Sigma_J = E\left\{\tfrac{1}{J}(\Sigma_{i=1}^{J}\underline{y}_i - \underline{\mu})(\Sigma_{i=1}^{J}\underline{y}_i - \underline{\mu})^{\dagger}\right\},$$

Since $\underline{b}_{MAI}(i)$ is independent of $\underline{b}_{MAI}(i+m)$ for $|m| \geq 3$, and $E\left\{\underline{b}_{MAI}(i)\right\} = 0$ for all i, using the assumption that the information bits of the interfering users are zero-mean, equally probable binary random variables, the above expression reduces to

$$\Sigma_J = S_{MAI} A_{MAI} \left(I_{K \times K} \otimes \frac{1}{J} \begin{bmatrix} J & J-1 & J-2 \\ J-1 & J & J-1 \\ J-2 & J-1 & J \end{bmatrix} \right) A_{MAI}^{\dagger} S_{MAI}^{\dagger} + \sigma^2 I_{N \times N}. \tag{5}$$

We note that $\Sigma = \lim_{J \to \infty} \Sigma_J$ is given by

$$\Sigma = \left[\sum_{k=1}^{K} \left(C_k \underline{\tilde{h}}_k(\underline{\tau}_k, \underline{\alpha}_k) \right) \left(C_k \underline{\tilde{h}}_k(\underline{\tau}_k, \underline{\alpha}_k) \right)^{\dagger} + \sigma^2 I_{N \times N} \right]. \tag{6}$$

Thus, the distribution of $\underline{m}_J$ can be approximated by a complex Gaussian distribution with covariance matrix $Q_J = \frac{1}{J}\Sigma_J$ and mean vector $\underline{\mu}$. Furthermore, for sufficiently large J, we can use the limiting distribution to replace the covariance matrix $Q_J = \frac{1}{J}\Sigma_J$ with its limit $Q = \frac{1}{J}\Sigma$, with Σ given above in (6). Thus, the asymptotic likelihood function for the parameter vector $\underline{\theta}_0 = [\underline{\tau}_0, \underline{\alpha}_0]$ is given as:

$$f_{\underline{m}_J}(\underline{m}_J; \underline{\theta}_0) \approx \frac{1}{\pi^N \det(Q)} e^{-(\underline{m}_J - C_0 \underline{\tilde{h}}_0(\underline{\tau}_0, \underline{\alpha}_0))^{\dagger} Q^{-1} (\underline{m}_J - C_0 \underline{\tilde{h}}_0(\underline{\tau}_0, \underline{\alpha}_0))}. \tag{7}$$

The use of the limiting distribution with covariance matrix Q is motivated by the fact that Q has a much simpler eigenstructure than Q_J. The signal subspace of Q is of rank K, whereas the signal subspace of Q_J is of rank p, where $2K \leq p \leq 3K$ depending the particular delays of the interfering users. The matrix Q in the estimation process will serve as a whitening filter to suppress the effects of MAI on the sample average $\underline{m}_J$. Our derivation of the Q matrix indicates that for sufficiently large J, the interfering users are only affecting $\underline{m}_J$ along K signal vectors. In the use of Q versus Q_J, the current work departs from prior work in this area.

It is patent from equation (6) that without explicit knowledge of the parameters of the interfering users, the covariance matrix Q is not explicitly known. Thus, we resort to estimation via relevant sample correlation matrices. The unbiased estimator of Q is,

$$\hat{Q} = \frac{1}{J-5} \left\{ \hat{Q}_{(i)(i)} + (\hat{Q}_{(i)(i+1)} + \hat{Q}^{\dagger}_{(i)(i+1)}) + (\hat{Q}_{(i)(i+2)} + \hat{Q}^{\dagger}_{(i)(i+2)}) - 5\underline{\hat{\mu}}\underline{\hat{\mu}}^{\dagger} \right\}, \tag{8}$$

where $\hat{Q}_{(i)(i+d)} = \frac{1}{J-d} \sum_{i=1}^{J-d} \underline{y}_i \underline{y}^{\dagger}_{i+d}$. Then the parameterization of Q given by (6) is employed to modify the technique of [3] to exploit the structure of the covariance matrix.

IV. ML Parameter Estimation

A. *Exact Maximum Likelihood*

The exact maximum likelihood estimator is the parameter vector that yields the maximum value of the likelihood function given in (7). The optimization problem expressed with the 2-norm is:

$$\begin{bmatrix} \underline{\hat{\alpha}}_0 \\ \underline{\hat{\tau}}_0 \end{bmatrix} = \arg \min_{\underline{\tau}_0, \underline{\alpha}_0} \| Q^{-1/2} \underline{m}_J - Q^{-1/2} C_0 D(\underline{\tau}_0) \underline{\alpha}_0 \|_2. \tag{9}$$

We observe that the problem is quadratic in $\underline{\alpha}_0$. Therefore, for a given set of delays $\underline{\tau}_0$, the $\underline{\alpha}_0$ which minimizes (9) is given by $\hat{\underline{\alpha}}_0 = (Q^{-1/2}C_0D(\underline{\tau}_0))^{\ddagger}Q^{-1/2}\underline{m}_J$, where $(\cdot)^{\ddagger}$ denotes the pseudo-inverse. Substituting the expression for $\hat{\underline{\alpha}}_0$ into (9) yields

$$\hat{\underline{\tau}}_0 = \arg\max_{\underline{\tau}_0} \|\mathbb{P}_{\tilde{D}(\underline{\tau}_0)}\tilde{\underline{m}}_J\|, \tag{10}$$

where $\tilde{D}(\underline{\tau}_0) = Q^{-1/2}C_0D(\underline{\tau}_0)$, $\tilde{\underline{m}}_J = Q^{-1/2}\underline{m}_J$ and $\mathbb{P}_{\tilde{D}(\underline{\tau}_0)}$ is the orthogonal projection to the subspace spanned by the columns of $\tilde{D}(\underline{\tau}_0)$.

Unfortunately, the objective function in (10) is not smooth because $D(\underline{\tau}_0)$ is discontinuous at chip boundaries. Therefore, we propose to consider two approximate ML techniques and compare their performance in Section VII via simulation studies.

B. *Approximate Maximum Likelihood via the EM Algorithm*

The expectation and maximization (EM) algorithm [4] can be used to decompose a large dimensional maximum likelihood estimation problem into tractable smaller ones. The estimation of the parameters $\underline{\alpha}_0$ and $\underline{\tau}_0$ from the observed data $\underline{m}_J$, is a maximum likelihood parameter estimation problem, where parameters of superimposed deterministic signals in colored Gaussian noise are searched. An approximate ML technique, which uses the EM algorithm, has been developed in [12] for the similar problem of estimating parameters of multiple sources in array signal processing. The EM algorithm for this multi-user parameter estimation application can be summarized as follows:

1. Set $p = 0$, initialize $\hat{\alpha}_{0,l}^{(0)}$ and $\hat{\tau}_{0,l}^{(0)}$.
2. E-step: For $l = 1 \ldots L$, compute $\hat{\underline{x}}_l^{(p)} = \hat{\alpha}_{0,l}^{(p)}\delta(\hat{\tau}_{0,l}^{(p)}) + \beta_l\left(C_0^{-1}\underline{m}_J - D(\hat{\underline{\tau}}_0{}^{(p)})\hat{\underline{\alpha}}_0{}^{(p)}\right)$.
3. M-step: For $l = 1 \ldots L$,

$$\begin{bmatrix} \hat{\alpha}_{0,l}^{(p+1)} \\ \hat{\tau}_{0,l}^{(p+1)} \end{bmatrix} = \arg\min_{\alpha,\tau} \left(\hat{\underline{x}}_l^{(p)} - \alpha\delta(\tau)\right)^{\dagger} C_0^T Q^{-1} C_0 \left(\hat{\underline{x}}_l^{(p)} - \alpha\delta(\tau)\right). \tag{11}$$

4. If the algorithm has not converged, set $p = p + 1$ and iterate steps 2-4.

The algorithm is said to have converged when the likelihood function has not increased, or has insignificantly increased from one iteration to the next. The initial estimate and the β_l's are design choices. In this study, the β_l is selected such that the above procedures result in a multi-stage-type estimation. That is, it is chosen such that $\beta_l = \delta(\tilde{p} - l)$ where $\tilde{p} = p \bmod L$. This choice provides a faster convergence rate for the algorithm as discussed in [13]. Then, the channel parameters associated with each propagation path can be alternately estimated, and all of the initial estimates can be set to zero.

The M-step of the algorithm has the highest computational cost. Closed form expressions for the M-step can be derived as

$$\hat{\tau}_{0,l}^{(p+1)} = \arg\max_{\tau} \frac{|\delta(\tau)^{\dagger}W\hat{\underline{x}}_l^{(p)}|^2}{\delta(\tau)^{\dagger}W\delta(\tau)} \quad \hat{\alpha}_{0,l}^{(p+1)} = \frac{\delta(\hat{\tau}_{0,l}^{(p+1)})^{\dagger}W\hat{\underline{x}}_l^{(p)}}{\delta(\hat{\tau}_{0,l}^{(p+1)})^{\dagger}W\delta(\hat{\tau}_{0,l}^{(p+1)})}, \tag{12}$$

where the weighting matrix W is defined as $W = C_0^TQ^{-1}C_0$. Since the function $\delta(\tau)$ is piecewise-continuous, the maximization can be achieved by first dividing the support of τ into N cells of chip period T_c in which $\delta(\tau)$ is continuous and differentiable, and finding the local maxima at each cell [2, 3, 14].

C. *Alternating Projections approximation to ML*

We can design an iterative procedure to solve for $\underline{\tau}_0$, using the projection structure of the objective function. The alternative projection algorithm is adopted from [5], where it was applied to the localization of multiple sources in passive sensor arrays.

The algorithm can be summarized as follows:

1. Set $p = 0$, initialize $\hat{\tau}_{0,l}^{(0)}$.
For $l = 1 \dots L$, repeat steps 2 and 3.
2. Projection Step: Compute the projection matrix $\mathbb{P}^{\perp}_{\tilde{D}_{-l}(\tau_{0,1}^{(p+1)},\dots,\tau_{0,l-1}^{(p+1)},\tau_{0,1+1}^{(p)},\dots,\tau_{0,L}^{(p)})}$,
where $\tilde{D}_{-l}$ is the matrix $\tilde{D}$ with the lth column deleted and $\mathbb{P}^{\perp}_{\tilde{D}_{-l}} = I - \mathbb{P}_{\tilde{D}_{-l}}$. We note that at each iteration step, the paths are estimated sequentially, not in parallel.
3. Maximization step:

$$\hat{\tau}_{0,l}^{(p+1)} = \arg\max_{\tau} \|\mathbb{P}_{(\mathbb{P}^{\perp}_{\tilde{D}_{-l}} Q^{-1/2} C_0 \delta(\tau))} \underline{\tilde{m}}_J \| x = \arg\max_{\tau} \frac{|\delta(\tau)^{\dagger} W \left(C_0^{-1} \underline{m}_J \right)|^2}{\delta(\tau)^{\dagger} W \delta(\tau)}, \tag{13}$$

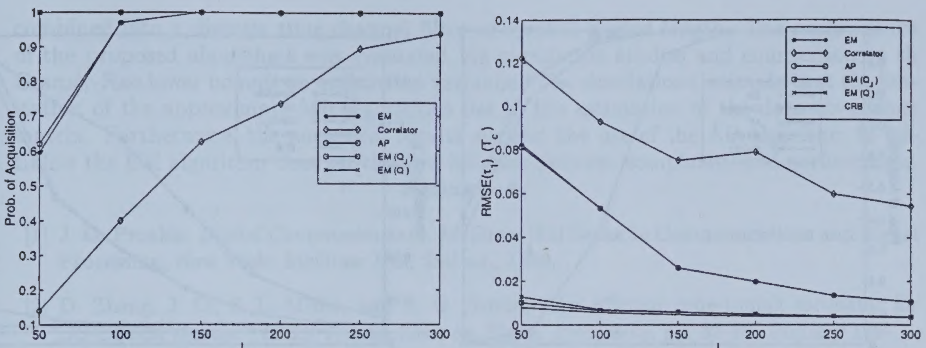
where W is given by $W = C_0^{\dagger} Q^{-1/2} \mathbb{P}^{\perp}_{\tilde{D}_{-l}} Q^{-1/2} C_0$.
4. If the algorithm has not converged, set $p = p + 1$ and iterate step 2-4.
5. Estimate amplitudes using: $\hat{\underline{\alpha}}_0 = (Q^{-1/2} C_0 D(\hat{\underline{\tau}}_O))^{\ddagger} Q^{-1/2} \underline{m}_J$.

Again the algorithm is said to have converged when the objective function has not increased, or has insignificantly increased from one iteration to the next. The maximization problem in (13) has the same form of the M-step in the EM algorithm and therefore the solution outlined in Section B. is applicable to this optimization problem as well.

V. Simulation Data

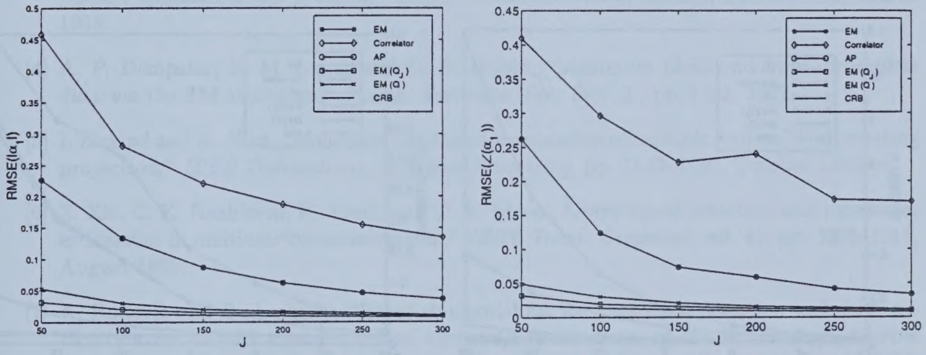
In the simulation experiments, all users are assigned Gold sequences of length $M = 31$ and multipath channels with $L = 3$ paths. The multipath delays are randomly chosen and fixed throughout the simulations. The support of the multipath delays is chosen from a uniform distribution $U[0, 3T/2)$ and the L paths are uniformly distributed within the support interval $(T/2)$. The phase for each path is chosen from a uniform distribution $U[0, 2\pi)$. In all simulations, the paths of the new user have unit power, *i.e.*, $|\alpha_{0,l}| = 1$. The signal to noise ratio is defined as $SNR = L/N\sigma^2$. The paths of the interfering users are assumed to have equal power $P_k = |\alpha_{k,l}|^2$ and multiple access interference is defined as $MAI = (P_k)$. In each example, the performance of estimators are measured using 250 Monte Carlo simulations. The following performance measures are used to quantify the performance of the different techniques: (a). Probability of correct acquisition is defined as the probability that all L paths are correctly acquired. A correct acquisition for a path l occurs, when the estimate is within a half chip of the true value, *i.e.*, $|\hat{\tau}_{0,l} - \tau_{0,l}| < T_c/2$. (b). RMSE($\hat{\tau}_{0,l}$) is the root mean square error of the delay estimate, given correct acquisition of path l (c) RMSE($|\hat{\alpha}_{0,l}|$) is the root mean square error of the magnitude of amplitude estimate, given correct acquisition of path l (d) RMSE($\angle\hat{\alpha}_{0,l}$) is the root mean square error of the phase of amplitude estimate, given correct acquisition of path l

In the first experiment, the number of active users is fixed at $K = 10$. An SNR level of 10dB and MAI level of 10dB is used. The length of the observation interval is varied between $J = 50$ and $J = 300$. The performance of the simple correlator [2], and the EM and AP based approximate ML estimators is given in Figures 2(a)–(d). The performance of the two EM based ML algorithms which assume perfect knowledge of the



(a) Prob. of Acquisition vs. J

(b) RMSE($\tau_{0,1}$) vs. J



(c) RMSE($|\alpha_{0,1}|$) vs. J

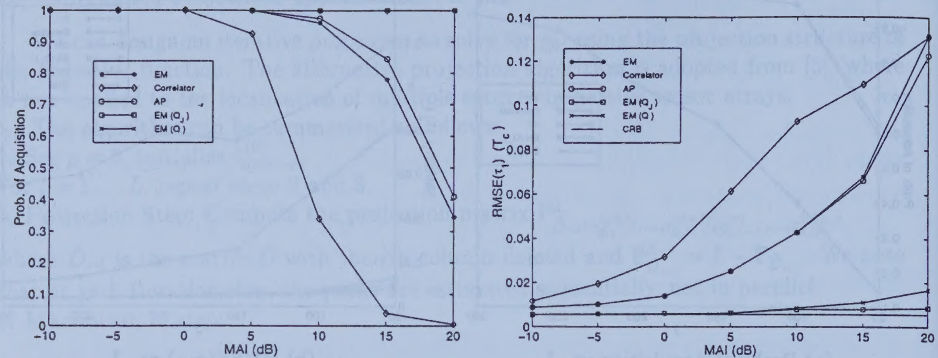
(d) RMSE($\angle(\alpha_{0,1})$) vs. J

Figure 2: SNR=10 dB, MAI=10 dB, K=10

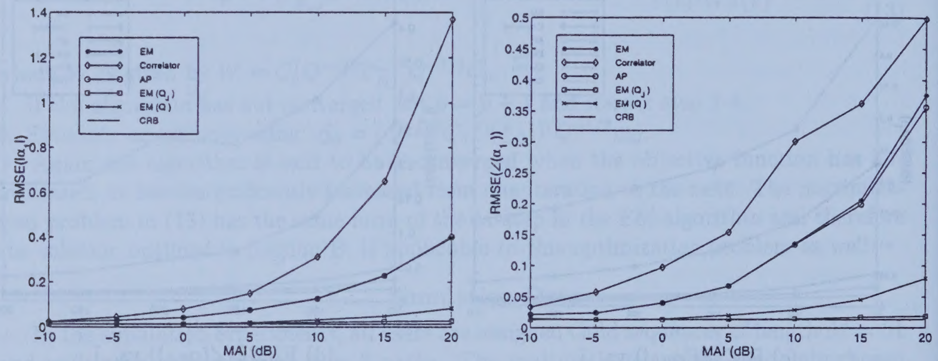
covariance matrix Q and Q_J, respectively, is also included. In Figure 2(a), the probability of acquisition is given for the two techniques for different values of J. In Figures 2(b)–(d), the RMSE of the delay and amplitude estimates is given, together with the Cramér-Rao bound of these parameters. For this experiment the performance of the EM algorithm with known covariance matrix Q_J is indistinguishable from the Cramér-Rao bound.

In the second experiment, the number of users is fixed at $K = 10$, an SNR level of 10dB and $J = 100$ bits is employed. The MAI is varied between -10dB to 20dB. The performance of the simple correlator, the EM and AP based approximate ML estimators is given in Figures 3(a)–d. In Figure 3(a), the probability of acquisition is given for the four techniques for different MAI levels. In Figures 3(b)–(d), the RMSE of the delay and amplitude estimates is given, together with the Cramér-Rao bound of these parameters. Again the performance of the EM algorithm with known covariance matrix Q_J is indistinguishable from the Cramér-Rao bound.

We first note that the performances of the two EM algorithms with perfect knowledge of Q and Q_J are identical, for a reasonable range of J and MAI values, which validates the use of the matrix Q in place of Q_J. Also the strong performance of these algorithms indicates that the suboptimal optimization provided by EM (or AP) is not the limiting factor of the proposed techniques. The performances of the EM and AP based techniques



(a) Prob. of Acquisition vs. MAI (b) RMSE($\tau_{0,1}$) vs. MAI



(c) RMSE($|\alpha_{0,1}|$) vs. J (d) RMSE($\angle(\alpha_{0,1})$) vs. MAI

Figure 3: SNR=10 dB, J=100, K=10.

are very similar. We have observed in our simulations that the computational cost of the EM is considerably smaller than that of AP, although it takes the AP algorithm less number of iterations to converge. These observations suggest the use of a hybrid technique, which uses the AP algorithm for initialization and the EM algorithm for additional iterations. The RMSE performance of the EM and AP algorithms approaches the Cramér-Rao bound for large observation intervals and the performance degrades for MAI levels higher than 20dB. Comparing the performance of the EM and AP algorithms with the techniques which use knowledge of the covariance matrix, we conclude that the poor performance of the EM and AP techniques at high MAI levels is due to the poor estimates of the covariance matrix Q and not due to the use of asymptotic form of the covariance matrix or due to the suboptimal optimization provided by EM and AP.

VI. Conclusion

We have proposed two iterative techniques for estimating the channel parameters of a new user in a multi-user multipath DS-CDMA communication environment. The proposed techniques are iterative implementations of the maximum likelihood estimators of the channel parameters under the assumption of as training sequence. We derived an asymptotic expression for the likelihood function of channel parameters and proposed a re-parametrization of this likelihood function, where the channel parameters are combined into a discrete time channel filter of symbol period length. The performance of the proposed algorithms was evaluated via simulation studies and comparison to th Cramér-Rao lower bounds on estimation variance. The simulations intimate that the limitation of the approximate ML algorithms lies in the estimation of the data covariance matrix. Furthermore, the numerical results suggest the use of the AP algorithm to initialize the EM algorithm thus providing a balance between complexity and performance.

References

[1] J. G. Proakis, *Digital Communications.* McGraw Hill Series in Communications and Signal Processing, New York: McGraw Hill, 2nd ed., 1989.

[2] D. Zheng, J. Li, S. L. Miller, and E. G. Strom, "An efficient code-timing estimator for DS-CDMA signals," *IEEE Transactions on Signal Processing*, pp. 82–89, January 1997.

[3] S. E. Bensley and B. Aazhang, "Maximum likelihood synchronization of a single user for CDMA communication systems," *IEEE Trans. Commun.*, vol. 46, pp. 392–399, March 1998.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J .Roy. Statistics. Soc. Ser. B.*, pp. 1–38, 1977.

[5] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Transactions on Signal Processing*, pp. 1553–1560, October 1988.

[6] Z. Xie, C. K. Rushforth, R. Short, and T. K. Moon, "Joint signal detection and parameter estimation in multiuser communications," *IEEE Trans. Commun.*, vol. 41, pp. 1208–1216, August 1993.

[7] A. Radovic and B. Aazhang, "Iterative algorithms for joint data detection and delay estimation for CDMA communication systems," in *in Proc. of the* 31^{st} *annual Allerton Conference on communication, control, and computing*, (Allerton House, Monticello, IL), January 1993.

[8] C. Sengupta, A. Hottinen, J. R. Cavallaro, and B. Aazhang, "Maximum likelihood multipath channel parameter estimation in CDMA systems." Rice University Technical Report, 1998.

[9] D. Sarwate and M. Pursley, "Properties of pseudo-random and related sequences," *Proceedings of the IEEE*, vol. 68, pp. 590–617, May 1980.

[10] G. Stuber, *Principles of Mobile Communication.* Boston, MA: Kluwer Academic Publishers, 1984.

[11] M. M. Rao, *Probability Theory with Applications*, vol. 1. Orlando, Florida: Academic Press, Inc., 1984.

[12] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 477–488, April 1988.

[13] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Transactions on Signal Processing*, pp. 2664–2677, October 1994.

[14] E. G. Strom, S. Parkvall, S. L. Miller, and B. E. Ottersten, "Propogation delay estimation in asynchronous direct sequence code-division multiple access systems," *IEEE Trans. Commun.*, vol. 44, pp. 84–93, January 1996.

Using Linear Prediction in Joint Blind Multiuser Detection and Blind Channel Equalization for CDMA Systems

H. Howard Fan* and Xiaohua Li
Department of ECECS, University of Cincinnati
Cincinnati, OH 45221-0030
Ph: (513) 556-4765
Fax: (513) 556-7326
E-mail: h.fan@uc.edu

Abstract

For high speed code division multiple access (CDMA) systems, interchip interference that is caused by multipath propagation becomes severe and can not be ignored. Together with the multiple access interference due to nonorthogonality in CDMA, they become major obstacles to overall system performance. In this paper we propose a new blind method based on linear prediction. It is known that linear prediction is robust in many ways, and is relatively simple in computation. When applied to the joint blind multiuser detection and blind equalization, it is suitable for both synchronous and asynchronous cases. In addition, it turns out that the new method is insensitive to error in propagation delay estimation, or timing recovery, in certain situations. Thus the proposed method is not only decorrelating and zero-forcing, it is also robust and near-far resistant.

1 Introduction

Many high speed digital communication systems suffer from the problem of intersymbol interference (ISI), which may arise from the common phenomenon of multipath propagation, for example. This includes high speed code division multiple access (CDMA) systems in which case it is the interchip interference (ICI) that results. Even for low speed CDMA systems, in indoor and urban environments ICI or ISI can not be ignored. To achieve reliable communication in these situations, channel equalization is necessary to eliminate ISI or ICI. In addition, multiuser detection and multiple access interference (MAI) rejection are simultaneously needed in a CDMA system. Furthermore, channel identification may also be needed in, for example, the RAKE receiver in a CDMA system. Therefore, joint multiuser detection and channel estimation/equalization are essential to practical CDMA systems.

Some existing work on joint equalization and multiuser detection methods are based on *a priori* knowledge of the channel [1], [2]. The blind methods, on the other hand, are much more desirable since they can eliminate MAI and ICI without requiring training or *a priori* knowledge. Blind multiuser detection methods have been proposed in [3] and [4] assuming no multipath or no ICI. Clearly they will not work properly when ICI can not be ignored.

---

* Corresponding author.

Very recently, several researchers have turned their attention to the joint blind multiuser detection and blind channel estimation/equalization problem. Two kinds of blind methods have been proposed. The first kind is subspace based methods [5] - [7]. These algorithms all require singular value decomposition (SVD) or eigenvalue decomposition (EVD) of some form of the data correlation matrix. The computational burdens in computing SVD or EVD are of order $O(m^3)$, where m is the dimension of the correlation matrix. Therefore, these algorithms are not suitable for real-time recursive implementation. Furthermore, it turns out that these SVD or EVD based algorithms are very sensitive to the rank estimation of the data correlation matrix. Our experience shows that the residual ISI or ICI may be significantly higher if the rank estimation is off even by one. In a practically noisy environment, accurate rank determination may be difficult. These two drawbacks may hinder wide-spread applications of these algorithms. The second kind is constrained optimization [8]. The algorithm is simple in computation. However, it turns out that the algorithm relies critically on accurate propagation delay estimation, which significantly increases the required computation. In addition, some performance loss in terms of minimum mean square error may result using this algorithm. This problem is due to the hard constraint used in the algorithm, which may be too restrictive in many cases.

In this paper, we propose a new blind method that is based on multichannel linear prediction. The new method is efficient in computation and robust in performance. Specifically, the computational complexity is at most in the order of $O((L_wL_c)^2)$ where L_wL_c is the equalizer (receiver) length. Furthermore, the new method is applicable to synchronous as well as asynchronous cases, and in the latter case the new method is insensitive to timing estimation error. It will be shown that the new method is both decorrelating and zero-forcing, and is also near-far resistant.

2 Problem Formulation

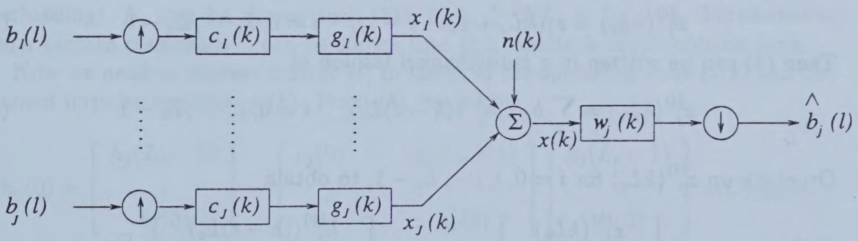


Figure 1: *CDMA receiver block-diagram*

Figure 1 is a block diagram of a discretized CDMA receiver for the jth user.

There are altogether J users in the entire system. $b_j(\ell)$ is the jth user's symbol sequence. $c_j(k)$ is the jth user's spreading code. $g_j(k)$ is the jth (multipath) channel impulse response. And $n(k)$ is the additive noise. Note that the received signal $x(k)$ has the same form for all users, which includes MAI (with ICI) due to other users and ICI due to the jth channel $g_j(k)$ for the jth user. For blind detection/equalization, $x(k)$ is the only signal that is available. Together with the known jth user's code sequence $c_j(k)$, our objective is to design the receiver for the jth user, $\mathbf{w}_j(k)$, that is decorrelating for MAI rejection and zero-forcing for ICI removal.

Let's consider the jth user. The chip "modulated" symbol stream can be expressed as

$$s_j(k) = \sum_{\ell} b_j(\ell) c_j(k - \ell L_c) \tag{1}$$

where $b_j(\ell)$ is at the symbol rate $\frac{1}{T_b}$, and $c_j(k)$ and $s_j(k)$ at the chip rate $\frac{1}{T_c}$, with L_c the length of the chip sequences. This signal is transmitted through the channel $g_{cj}(t)$ which is assumed to be causal. The jth component of the received signal is then

$$x_{cj}(t) = \sum_{k} \sum_{\ell} b_j(\ell) c_j(k - \ell L_c) g_{cj}(t - kT_c - \tau_j). \tag{2}$$

Assuming τ_j is known through timing recovery (later we will discuss ways to estimate τ_j blindly) and sampling at $t = kT_c + \tau_j$, we have

$$x_j(k) = \sum_{i} \sum_{\ell} b_j(\ell) c_j(i - \ell L_c) g_j(k - i) \tag{3}$$

where the notation $g_j(k) = g_{cj}(kT_c)$ and $x_j(k) = x_{cj}(kT_c + \tau_j)$ are used. It is straight forward to show that (3) is equivalent to

$$x_j(k) = \sum_{\ell} b_j(\ell) h_j(k - \ell L_c) \tag{4}$$

$$h_j(k) = \sum_{\ell=0}^{L_c-1} c_j(\ell) g_j(k - \ell) \tag{5}$$

Define

$$h_j^{(i)}(kL_c) = h_j(kL_c + L_c - i - 1), \quad i = 0, 1, \cdots, L_c - 1 \tag{6}$$

$$x_j^{(i)}(kL_c) = x_j(kL_c + L_c - i - 1), \quad i = 0, 1, \cdots, L_c - 1 \tag{7}$$

Then (4) can be written in a multichannel fashion as

$$x_j^{(i)}(kL_c) = \sum_{\ell} b_j(\ell) h_j^{(i)}((k - \ell)L_c), \quad i = 0, 1, \cdots, L_c - 1 \tag{8}$$

Or, stack up $x_j^{(i)}(kL_c)$ for $i = 0, 1, \cdots, L_c - 1$, to obtain

$$\begin{bmatrix} x_j^{(0)}(kL_c) \\ x_j^{(1)}(kL_c) \\ \vdots \\ x_j^{(L_c-1)}(kL_c) \end{bmatrix} = \sum_{\ell} b_j(\ell) \begin{bmatrix} h_j^{(0)}((k-\ell)L_c) \\ h_j^{(1)}((k-\ell)L_c) \\ \vdots \\ h_j^{(L_c-1)}((k-\ell)L_c) \end{bmatrix} \tag{9}$$

With obvious vector notations, (9) can be written as

$$\begin{aligned}\mathbf{x}_j(k) &= \sum_{\ell} b_j(\ell)\mathbf{h}_j(k-\ell)\\ &= \sum_{\ell=0}^{L_h-1} b_j(k-\ell)\mathbf{h}_j(\ell) \qquad (10)\end{aligned}$$

Now furthermore stacking up $\mathbf{x}_j(k), \mathbf{x}_j(k-1), \cdots, \mathbf{x}_j(k-L_w+1)$ and using (10), we have

$$\begin{bmatrix} \mathbf{x}_j(k) \\ \mathbf{x}_j(k-1) \\ \vdots \\ \mathbf{x}_j(k-L_w+1) \end{bmatrix} = \begin{bmatrix} \mathbf{h}_j(0) & \cdots & \mathbf{h}_j(L_h-1) & & \mathbf{0} \\ & \ddots & & \ddots & \\ & & \ddots & & \ddots \\ \mathbf{0} & & & \mathbf{h}_j(0) & \cdots \quad \mathbf{h}_j(L_h-1) \end{bmatrix} \cdot \begin{bmatrix} b_j(k) \\ b_j(k-1) \\ \vdots \\ b_j(k-L_h-L_w+2) \end{bmatrix} \qquad (11)$$

where L_h is related to the length of $h_j(k)$, and L_w is related to the length of $\mathbf{w}_j(k)$. Or, again with obvious vector and matrix notations, (11) can be written as

$$\mathcal{X}_j(k) = \mathcal{H}_j \mathbf{b}_j(k) \qquad (12)$$

Ignoring the noise $n(k)$ for the time being, the received signal vector $\mathcal{X}(k) = [x(kL_c + L_c - 1), x(kL_c + L_c - 2), \cdots, x(kL_c), x(kL_c - 1), \cdots, x(kL_c - L), \cdots, x(kL_c - (L_w - 1)L_c + L_c - 1), \cdots, x(kL_c - (L_w - 1)L_c)]^T$ can then be expressed as, using (12),

$$\mathcal{X}(k) = \sum_{j=1}^{J} \mathcal{X}_j(k) = \begin{bmatrix} \mathcal{H}_1 & \cdots & \mathcal{H}_J \end{bmatrix} \begin{bmatrix} \mathbf{b}_1(k) \\ \vdots \\ \mathbf{b}_J(k) \end{bmatrix} \qquad (13)$$

The matrix $[\mathcal{H}_1 \ \cdots \ \mathcal{H}_J]$ is of dimension $L_w L_c \times J(L_h + L_w - 1)$. For $L_c > J$ (no overloading), L_w can be chosen such that $L_w L_c \geq J(L_h + L_w - 1)$. Furthermore, under certain conditions it can be shown that this matrix is of full column rank.

Now we need to express matrix $\mathcal{H}_j$ in terms of the spreading code $c_j(k)$ and the channel impulse response $g_j(k)$. From (5), we have

$$\mathbf{h}_j(0) = \begin{bmatrix} h_j(L_c-1) \\ \vdots \\ h_j(0) \end{bmatrix} = \begin{bmatrix} c_j(0) & \cdots & c_j(L_c-1) \\ & \ddots & \vdots \\ \mathbf{0} & & c_j(0) \end{bmatrix} \begin{bmatrix} g_j(L_c-1) \\ \vdots \\ g_j(0) \end{bmatrix} = C_{j0}\mathbf{g}_j(0) \qquad (14)$$

In a cellular CDMA environment the maximum delay spread of the channel is usually much less than the symbol duration, i.e., the channel length is much smaller than the spreading ratio L_c [9]. In other words, the vector $\mathbf{g}_j(0)$ is usually not full. Therefore it is the only vector describing the jth channel. As a consequence, $L_h = 2$, and we have from (5),

$$\mathbf{h}_j(1) = \begin{bmatrix} h_j(2L_c-1) \\ \vdots \\ h_j(L_c) \end{bmatrix} = \begin{bmatrix} 0 & \cdots\cdots & \cdots & 0 \\ c_j(L_c-1) & & & \vdots \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ c_j(1) & \cdots\cdots & c_j(L_c-1) & 0 \end{bmatrix} \mathbf{g}_j(0) = C_{j1}\mathbf{g}_j(0) \tag{15}$$

The matrix $\mathcal{H}_j$ then becomes

$$\begin{bmatrix} \mathbf{h}_j(0) & \mathbf{h}_j(1) & & 0 \\ & \ddots & \ddots & \\ 0 & & \mathbf{h}_j(0) & \mathbf{h}_j(1) \end{bmatrix} = \begin{bmatrix} C_{j0} & C_{j1} & & 0 & \vdots & 0 \\ & \ddots & \ddots & & \vdots & \vdots \\ & & \ddots & C_{j1} & \vdots & 0 \\ 0 & & & C_{j0} & \vdots & C_{j1} \end{bmatrix} \begin{bmatrix} \mathbf{g}_j(0) & & 0 \\ & \ddots & \\ 0 & & \mathbf{g}_j(0) \end{bmatrix} \tag{16}$$

or, with obvious matrix notations

$$\mathcal{H}_j = \begin{bmatrix} \mathcal{C}_j & \vdots & \begin{matrix} 0 \\ C_{j1} \end{matrix} \end{bmatrix} \mathcal{G}_j \tag{17}$$

The matrix $\mathcal{C}_j$ in (17) is non-singular if $c_j(0) \neq 0$. Therefore $\mathcal{C}_j^{-1}$ exists and can be easily computed beforehand. Now consider $\mathcal{C}_j^{-1}\mathcal{H}_j$. Using (17) we have

$$\mathcal{C}_j^{-1}\mathcal{H}_j = \begin{bmatrix} & \vdots & * \\ I & \vdots & \vdots \\ & \vdots & * \end{bmatrix} \mathcal{G}_j = \begin{bmatrix} \mathbf{g}_j(0) & \mathbf{0} & & \bigcirc & * \\ & \mathbf{g}_j(0) & & & \vdots \\ & & \ddots & & \vdots \\ \bigcirc & & & \mathbf{g}_j(0) & * \end{bmatrix} \tag{18}$$

Note that the only non-zero components in the second column is $\mathbf{g}_j(0)$ which is in the second block of rows. On the other hand, in $\mathcal{C}_j^{-1}\mathcal{H}_i$ for $i \neq j$, since $\mathcal{C}_j^{-1}\mathcal{C}_i$ is upper triangular and is not identity, we have

$$\mathcal{C}_j^{-1}\mathcal{H}_i = \begin{bmatrix} * & \cdots & \cdots & * \\ & \ddots & & \vdots \\ \bigcirc & & * & * \end{bmatrix} \mathcal{G}_i = \begin{bmatrix} * & * & \cdots & * & * \\ & * & & \vdots & \vdots \\ & & \ddots & \vdots & \vdots \\ \bigcirc & & & * & * \end{bmatrix} \tag{19}$$

That is, there is no column with an only non-zero block in the second block of rows. This is the key in using linear prediction.

3 Linear Prediction Approach

Since $c_j(k)$ is known to the jth user, we can multiply C_j^{-1} to the received signal vector $\mathcal{X}(k)$. Using (13), (18), and (19), we have

$$C_j^{-1}\mathcal{X}(k) = \left[\begin{bmatrix} * & * & \cdots & * & * \\ & * & & \vdots & \vdots \\ & & \ddots & \vdots & \vdots \\ \bigcirc & & & * & * \end{bmatrix} \cdots \begin{bmatrix} \mathbf{g}_j(0) & \mathbf{0} & & \bigcirc & * \\ & \mathbf{g}_j(0) & & & \vdots \\ & & \ddots & & \vdots \\ \bigcirc & & & \mathbf{g}_j(0) & * \end{bmatrix} \cdots \begin{bmatrix} * & * & \cdots & * & * \\ & * & & \vdots & \vdots \\ & & \ddots & \vdots & \vdots \\ \bigcirc & & & * & * \end{bmatrix} \right] \begin{bmatrix} \mathbf{b}_1(k) \\ \vdots \\ \mathbf{b}_J(k) \end{bmatrix} \tag{20}$$

or

$$C_j^{-1}\mathcal{X}(k) = Q \begin{bmatrix} \mathbf{b}_1(k) \\ \vdots \\ \mathbf{b}_J(k) \end{bmatrix} \tag{21}$$

Decompose Q as follows

$$Q = \begin{bmatrix} Q_1 \\ Q_0 \\ Q_2 \end{bmatrix} \tag{22}$$

where Q_0 and Q_1 each has L_c rows, respectively. Define an L_c dimensional prediction error vector

$$\varepsilon(k) = \begin{bmatrix} -P_1 & I & -P_2 \end{bmatrix} C_j^{-1}\mathcal{X}(k) \tag{23}$$

where P_1 is $L_c \times L_c$. Assuming that the symbols $b_j(k)$ are uncorrelated in time and that $\mathbf{b}_1(k), \cdots, \mathbf{b}_J(k)$ are mutually uncorrelated with variances (powers) $A_1, \cdots, A_J$. Then it can be seen, using (21)-(23),

$$\begin{aligned} E\{\varepsilon(k)\varepsilon^H(k)\} &= \begin{bmatrix} -P_1 & I & -P_2 \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_0 \\ Q_2 \end{bmatrix} A \begin{bmatrix} Q_1^H & Q_0^H & Q_2^H \end{bmatrix} \begin{bmatrix} -P_1^H \\ I \\ -P_2^H \end{bmatrix} \\ &= [Q_0 - P_1Q_1 - P_2Q_2]A[Q_0^H - Q_1^H P_1^H - Q_2^H P_2^H] \end{aligned} \tag{24}$$

where $A = \mathrm{diag}\{A_1 I, \cdots, A_J I\}$. Minimizing $E\{\varepsilon(k)\varepsilon^H(k)\}$ over P_1 and P_2 then gives

$$\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} A \left(Q_0 - \begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}\right)^H = \bigcirc. \tag{25}$$

Due to the structure of (20), the matrix $\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} A$ has one and only one zero column at the second column of the jth block. Therefore,

$$Q_0 - \begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = \begin{bmatrix} -P_1 & I & -P_2 \end{bmatrix} \begin{bmatrix} C_j^{-1}\mathcal{H}_1 & \cdots & C_j^{-1}\mathcal{H} \end{bmatrix} \tag{26}$$

has only one non-zero column also at the second column of the jth block. But this column of $\begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$ is zero since this column of $\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$ is zero. Therefore this non-zero column of (26) is nothing but the second column of the jth block of Q_0, i.e., $\mathbf{g}_j(0)$. In other words,

$$\begin{bmatrix} -P_1 & I & -P_2 \end{bmatrix} \begin{bmatrix} \mathcal{C}_j^{-1}\mathcal{H}_1 & \cdots & \mathcal{C}_j^{-1}\mathcal{H}_J \end{bmatrix} = [\, \mathbf{0} \cdots \mathbf{0} \; \underbrace{\mathbf{0} \; \mathbf{g}_j(0) \; \mathbf{0} \cdots \mathbf{0}}_{j\text{th block}} \; \mathbf{0} \cdots \mathbf{0} \,] \tag{27}$$

which yields

$$\min_{P_1, P_2} E\{\varepsilon(k)\varepsilon^H(k)\} = A_j \mathbf{g}_j(0)\mathbf{g}_j^H(0) \tag{28}$$

In other words, any row or column of (28) is proportional to $\mathbf{g}_j(0)$.

Now reconsider (13). Using (21), (22), (26), and (27), we have

$$\begin{aligned} \begin{bmatrix} -P_1 & I & -P_2 \end{bmatrix} \mathcal{C}_j^{-1}\mathcal{X}(k) &= \begin{bmatrix} -P_1 & I & -P_2 \end{bmatrix} \begin{bmatrix} \mathcal{C}_j^{-1}\mathcal{H}_1 & \cdots & \mathcal{C}_j^{-1}\mathcal{H}_J \end{bmatrix} \begin{bmatrix} \mathbf{b}_1(k) \\ \vdots \\ \mathbf{b}_J(k) \end{bmatrix} \\ &= b_j(k-1)\mathbf{g}_j(0) \end{aligned} \tag{29}$$

That is, the optimum equalizer/receiver vector should be

$$\mathbf{w}_j^H(k) = \mathbf{g}_j^H(0) \begin{bmatrix} -P_1 & I & -P_2 \end{bmatrix} \mathcal{C}_j^{-1} \tag{30}$$

which yields, when applied to $\mathcal{X}(k)$

$$\mathbf{w}_j^H(k)\mathcal{X}(k) = \|\mathbf{g}_j(0)\|^2 b_j(k-1) \tag{31}$$

Or in other words, other than a scaling factor, the jth user's symbol is estimated by

$$\hat{b}_j(k-1) = \mathbf{w}_j^H(k)\mathcal{X}(k) = \mathbf{g}_j^H(0)\varepsilon(k) \tag{32}$$

As to the implementation, the multichannel linear prediction (23) and (28) can be done using an on-line recursive gradient search algorithm. Only matrix-to-vector multiplication and additions are needed. The overall computation per iteration is therefore in the order of $O((L_w L_c)^2)$. Timing recovery can also be done by varying the number of rows of Q_1 from 1 to L_c and choosing an optimum one. However, in case $\mathbf{g}_j(0)$ is not filled, such as in a cellular CDMA environment, a small amount of delay only shifts the non-zero components within $\mathbf{g}_j(0)$ and does not affect the performance of the algorithm. See [10] for more details.

References

[1] Z. Zvonar and D. Brady, "Optimum detection in asynchronous multiple-access multipath Rayleigh fading channels," in *Proc. 26th Conf. Infor. Sci. Systems*, March 1992.

[2] M. K. Varanasi and S. Vasudevan, "Multiuser detectors for synchronous CDMA communication over nonselective Rician fading channels," *IEEE Trans. Commun.*, Vol. 42, pp.711-722, Feb. 1994.

[3] M. Honig, U. Madhow, and S. Verdú, "Blind multiuser detection," *IEEE Trans. on Information Theory*, Vol. 41, pp. 944-960, July 1995.

[4] X. Wang and H. V. Poor, "Blind multiuser detection: A subspace approach," *IEEE Trans. Information Theory*,

[5] S. E. Bensley and B. Aazhang, "Subspace-based channel estimation for code division multiple access communication systems," *IEEE Trans. on Communications*, Vol. 44, No. 8, pp. 1009 - 1020, August 1996.

[6] M. Torlak and G. Xu, "Blind multiuser channel estimation in asynchronous CDMA systems," *IEEE Trans. on Signal Processing*, Vol. 45, No. 1, pp. 137 - 147, Jan. 1997.

[7] X. Wang and H. V. Poor, "Blind equalization and multiuser detection in dispersive CDMA channels," *IEEE Trans. on Communications*, vol. 46, no. 1, pp. 91 - 103, January 1998.

[8] M. K. Tsatsanis, "Inverse filtering criteria for CDMA systems," *IEEE Trans. on Signal Processing*, vol. 45, no. 1, pp. 102 - 112, January 1997.

[9] W. C. Y. Lee, "Overview of cellular CDMA," *IEEE Trans. Veh. Technol.*, pp. 291-301, May 1991.

[10] H. Fan and X. Li, "Joint blind multiuser detection and blind channel equalization in CDMA by linear prediction," submitted to ICC'99.

On the Dimensional Limitations of Linear Multiuser Detection*

Sridhar Gollamudi and Yih-Fang Huang
Dept. of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556.
huang.2@nd.edu
http://www.nd.edu/~aspect/

Abstract

Limitations on the interference suppression capabilities of linear multiuser detection due to dimensional constraints are studied in this paper. Generalized linear multiuser detection is defined for CDMA systems with arbitrary modulation and coding schemes and channel characteristics. Conditions are derived for the existence of linear decorrelators in this general framework and a solution for the generalized linear decorrelator is presented when it exists. This theory is applied to determine the efficacy of linear multiuser detection in important scenarios that include the presence of user asynchronism, multipath propagation, error control coding and block modulation (such as M-ary orthogonal modulation in IS-95 uplink). Ways to overcome the dimensional limitations of linear multiuser detection are to use larger symbol constellations that span a smaller number of dimensions, or to employ nonlinear techniques such as nonlinear multistage interference cancellation.

1 Introduction

The class of *linear* multiuser detection techniques for CDMA systems has attracted widespread attention in the past decade [1, 2, 3, 4], mainly due to their simplicity, amenability to adaptive implementations, and analytical tractability. Linear detectors for BPSK-modulated systems compute soft estimates of the transmitted bits as linear functions of the received signal, which are then hardlimited to make bit decisions. Since linear techniques operate by linearly separating groups of signals in the signal space, considerations on the dimensionalities of spaces occupied by desired and interfering signals are crucial in determining whether such techniques are effective in a given situation.

A generalized framework for linear detection of direct-sequence CDMA signals in the presence of multiuser interference is developed in this paper. This framework for linear detection is applicable in a variety of important scenarios such as asynchronous reception, multipath propagation, and the presence of channel coding or block modulation. Fundamental limitations of linear detection imposed by dimensional constraints in several realistic scenarios are delineated using this general framework. In this paper, linear

---

*This work was supported by the National Science Foundation under Grant MIP 9705173.

detection is considered *effective* in a given scenario if a generalized linear decorrelator (GLD) exists. This analysis serves to provide a better understanding of the effectiveness of linear detectors in practical CDMA systems.

In the case of a BPSK modulated system in the presence of user asynchronism and multipath propagation, this paper shows that a linear decorrelating receiver that operates on the received signal over a *finite* interval exists if the number of users is smaller than the spreading gain. This dispels a popular notion (see, e.g., [5]) that multipath propagation increases the number of dimensions of each user's signals and consequently reduces the number of interfering signals that can be suppressed using linear multiuser detectors. However, this paper also shows that the observation interval over which the linear detector must process the received signal in order to make a decision increases rapidly with increase in multipath delay spread. If large observation intervals cannot be tolerated (due to delay, complexity or adaptivity considerations), we show that asynchronism and/or multipath propagation effectively decreases the maximum allowable users with linear reception by a factor of two. Also, whenever the number of users is smaller than half the spreading gain, it is shown that GLD's that operate over a single bit interval (*one-shot* decorrelators) exist, regardless of asynchronism and the lengths of multipath delay spreads.

This paper shows that linear decorrelating detection is possible in the presence of "linear" (or single dimensional) modulation, "non-linear" (or multi-dimensional) modulation and channel coding. The dimension of the space spanned by the code symbols is shown to determine the interference suppression capability of linear detectors. In cases where the dimension of the space spanned by BPSK modulated codewords is smaller than the number of encoded bits in each codeword (*rank deficient codebooks*), GLD's derived in this paper offer more interference suppression capability compared to the conventional linear decorrelators that do not utilize the knowledge of the code. It is shown in this paper that *with block orthogonal coding (such as M-ary orthogonal modulation used on IS-95 uplink), the number of interferers that can be suppressed by a linear detector decreases exponentially with increase in block length.*

Linear multiuser detection schemes have been suggested in literature for use with nonlinear modulation schemes (e.g. the decision projection algorithm in [6]), which are also subject to the limitations derived in this paper. In particular, the maximum number of interferers that can be suppressed using such algorithms in an M-ary orthogonal modulated system also decreases exponentially with block length. Also, the so-called "soft-decision" interference cancelling receivers, which estimate the transmitted signal of each user as a linear function of the matched filter output for cancellation [7, 8], are also linear receivers and hence are subject to the constraints presented here.

Results from this paper indicate that the number of interferers that a linear detector can suppress can be increased beyond the spreading gain if a modulation scheme is used in which the symbol waveforms span a number of dimensions smaller than the number of bits per symbol. Examples of such schemes are PAM and M-PSK. Another way to overcome the dimensional constraints of linear detection is to employ nonlinear receivers. A number of attractive nonlinear multiuser receivers such as nonlinear successive or parallel interference cancelers and decision-feedback multiuser detectors are known. Nonlinear techniques are attractive also because most of them do not require, unlike adaptive linear multiuser detectors, the spreading codes to repeat every symbol (*short codes*).

The organization of this paper is as follows. Section 2 provides a general framework for linear multiuser detection and defines a class of generalized linear decorrelating detectors. Also, in Section 2, conditions for the existence of GLD's are derived, and solutions are provided when they do exist. These concepts are used in Section 3 to derive GLD's and study the limitations of linear detection in the presence of asynchronism and multipath propagation. Section 4 analyzes how channel coding and block modulation constrain the effectiveness of linear multiuser detection, and presents GLD's (when they exist) for coded systems.

2 Generalized Linear Multiuser Detection

2.1 Preliminaries

Consider K users simultaneously transmitting signals to a central receiver in a direct-sequence CDMA system. The received signal can be represented in complex baseband notation as

$$r(t) = s_1(t) + s_2(t) + \cdots + s_K(t) + n(t), \tag{1}$$

where $s_k(t)$ is the signal received from the kth user ($k = 1, 2, \cdots, K$), and $n(t)$ is additive noise. The kth user's received signal is obtained from the following multipath model:

$$s_k(t) = A_{k1}x_k(t - \tau_{k1}) + A_{k2}x_k(t - \tau_{k2}) + \cdots + A_{kJ}x_k(t - \tau_{kJ}), \tag{2}$$

where A_{kj} and τ_{kj} are the complex-valued channel gain and delay, respectively, on the jth path. Assume, without loss of generality, that $\tau_{11} = 0$. The kth user's transmitted signal, $x_k(t)$, is a product of kth user's modulated symbol waveforms and the spreading code of user k. Let T_b denote the bit-interval in the information bit-stream of any user. The modulation and coding schemes determine the dependence of modulated symbols on the information bits of the kth user. The spreading code consists of a sequence of *chip pulses*, each spanning T_c seconds. Spreading gain N of the CDMA system is defined as the number of chip pulses in one bit interval: $N = T_b/T_c$.

Without loss of generality, our attention will be limited to detecting user 1's signals. Let $[0, T]$, where T is an arbitrary integer multiple of T_c, be a finite observation interval such that a linear multiuser detector acts on $\{r(t) : t \in [0, T]\}$ to detect the desired user's signal in that interval. Note that this interval can span several bits of the desired user: we are not restricted to *one-shot* multiuser detectors. We shall be concerned in this paper only with linear detectors that span finite intervals ($T < \infty$). It is shown in Section 3.3 that a GLD for $T < \infty$ exists if *any* GLD exists.

The received signal in the observation interval $[0, T]$ is chip-matched filtered (synchronized to the delay of user 1) to obtain a vector of chip-rate samples, $\mathbf{r}$, prior to further processing[1]. The length of $\mathbf{r}$ is equal to the number of chips in the observation interval, $N_c = T/T_c$. Let $\mathcal{M}(\cdot)$ denote the chip-matched filtering operation that linearly maps a received signal in $[0, T]$ to the corresponding N_c-tuple of chip-matched filter samples. The following concepts are required to define generalized linear multiuser detection.

$\mathcal{S}_D$: Desired signal space, spanned by all possible desired user's signal vectors at the output of chip-matched filter (in the absence of interference). If $\mathcal{D}$ is the set of all possible user 1's signals over $t \in [0, T]$, then $\mathcal{S}_D = \text{Span}\{\mathcal{M}(s_t) : s_t \in \mathcal{D}\}$. It is determined by the symbol constellation, the choice of signature codes for the desired user, and channel characteristics. Let N_D denote its dimension, and let the columns of $B_D \in \mathbb{C}^{N_c \times N_D}$ constitute a basis[2]. *Example:* In an uncoded, single path BPSK system with $T = T_b$, we have $N_D = 1$, and $B_D = \mathbf{c}_1$, where $\mathbf{c}_1$ is the spreading code vector of user 1.

---

[1] For simplicity, it is assumed that the path delays of the desired user (τ_{1j}, $j = 1, 2, \cdots, J$) are integer multiples of the chip interval T_c, so that chip-matched filtering does not remove multipath information.

[2] $\mathbb{C}$ denotes the complex field.

$\mathcal{S}_I$: Interference space, spanned by the set of chip-matched filter output vectors corresponding to all possible interfering signals. In other words, if $\mathcal{I}$ is the set of all possible interfering signals over $t \in [0, T]$, then $\mathcal{S}_I = \text{Span}\{\mathcal{M}(i_t) : i_t \in \mathcal{I}\}$. This admits the possibility of chip-asynchronous interferers, since only the image of the interfering signals at the output of the desired user's chip-matched filter is of concern. Let N_I be the dimension of $\mathcal{S}_I$, and let $B_I \in \mathbb{C}^{N_c \times N_I}$ comprise a basis for it. *Example:* In a synchronous, single path CDMA system with uncoded BPSK modulation and $T = T_b$, we have $N_I = K - 1$, and $B_I = [\mathbf{c}_2 \; \mathbf{c}_3 \; \cdots \mathbf{c}_K]$, if all spreading codes are assumed to be linearly independent.

In terms of the above definitions, the received signal vector can be written as

$$\mathbf{r} = B_D \mathbf{u} + B_I \mathbf{w} + \mathbf{n}, \tag{3}$$

where $\mathbf{u} \in \mathbb{C}^{N_D}$ uniquely determines the transmitted signal of user 1, $B_I \mathbf{w}$ is the total interference signal at the output of the chip-matched filter, and $\mathbf{n} = \mathcal{M}(n(t); t \in [0, T])$. A *generalized linear multiuser detector* (GLMD) is defined by a matrix $V \in \mathbb{C}^{N_c \times N_D}$ that describes a linear mapping of $\mathbf{r}$ to an estimate of $\mathbf{u}$,

$$\hat{\mathbf{u}} = V^T \mathbf{r}, \tag{4}$$

followed by a minimum-distance detector that maps $\hat{\mathbf{u}}$ to a decision on the desired user's transmitted signal (without any knowledge of the interferers' codes). The above definition of a GLMD is extremely general, and coincides with the conventional definition of a linear multiuser detector in the case of BPSK modulated, single-path channel CDMA systems with no coding and $T = T_b$. The following sections elaborate on how this framework permits linear detection of signals in several other important scenarios.

In most prior works on linear multiuser detection, coded BPSK systems have been treated as uncoded systems by assuming that the linear detector has no knowledge of the coding scheme. In the following sections, an uncoded system refers to a system with no channel code, or a coded system in which the linear receiver estimates coded bits, as opposed to information bits using knowledge of the code. Separating linear multiuser detection and decoding operations at the receiver leads to the coding-spreading tradeoff, which is discussed further in Section 4.

2.2 The generalized linear decorrelator

Among linear multiuser detectors, of particular concern in the absence of noise is the linear decorrelating detector. A linear decorrelating detector is usually defined (see, e.g., [9, 10]) for uncoded BPSK systems with single path channels as a set of linear detectors for all K users, whose soft estimates equal the transmitted bits in the absence of noise.

In the framework of GLMD, a *generalized linear decorrelator* (GLD) for user 1 can be defined as a GLMD (cf. (4)) such that $\hat{\mathbf{u}} = \mathbf{u}$, for all possible received signals with $\mathbf{n} = 0$. In other words, the columns of V belong to the null space of $\mathcal{S}_I$, and are specified as follows[3]:

$$\begin{aligned} V^T B_D &= I_{N_D \times N_D}, & (5)\\ V^T B_I &= 0_{N_D \times N_I}. & (6) \end{aligned}$$

---

[3] $I_{N \times N}$ denotes the $N \times N$ identity matrix, and $0_{N \times M}$ denotes the $N \times M$ zero matrix.

In the case of uncoded, single path BPSK, $B_D = \mathbf{c}_1$ and $\mathbf{u} = \pm 1$; therefore, the GLD specification reduces to $\mathbf{v}^T\mathbf{c}_1 = 1$ and $\mathbf{v}^T B_I = 0_{1\times N_I}$. This coincides with the conventional definition of a linear decorrelator.

The following result can be inferred from the definition of the GLD. Its proof is omitted due to lack of space.

Proposition 2.1: A necessary condition for the existence of a GLD as defined in (5) and (6) is that

$$N_D + N_I \leq N_c, \tag{7}$$

and a sufficient condition for its existence is that $B = [B_D \; B_I]$ has full column rank.

The above sufficient condition is similar to the linear independence assumption frequently used in literature on conventional linear multiuser detection. For synchronous, uncoded, single path BPSK systems with $T = T_b$, we have $N_D = 1$, $N_I = K - 1$, and $N_c = N$. Therefore, the above necessary condition reduces to the familiar dimensional limitation of the linear decorrelator: $K \leq N$.

In the remainder of the paper, we shall assume that the transmitted signals from different users are linearly independent if (7) is satisfied, so that the sufficient condition in Proposition 2.1 is satisfied whenever (7) is. This is similar to making the linear independence assumption for a conventional decorrelator. If B has full column rank, the solution set for each column of V from (5) and (6) is a subspace with $N_c - (N_D + N_I)$ dimensions. Any choice of vectors from these subspaces forms a valid GLD, but the solution in which each column of V has minimum 2-norm is given by

$$V = B(B^T B)^{-1} \begin{bmatrix} I_{N_D\times N_D} \\ 0_{N_D\times N_I} \end{bmatrix}. \tag{8}$$

Minimizing the 2-norm of the columns of V minimizes the response of the GLD to additive noise.

3 Effect of Asynchronism and Multipath Propagation

This section investigates the existence of a GLD in the presence of asynchronous reception and multipath channels. Since the GLD is the most general form of linear decorrelating receiver, the results obtained in this section characterize the fundamental dimensional limitations of linear reception in such situations.

3.1 Asynchronous users, single path channels

Consider asynchronous reception of the K users' uncoded BPSK signals through single path channels. Let the observation interval $[0, T]$ span exactly M bits of user 1, as shown in Figure 1. It follows that the number of dimensions of the desired signal space is $N_D = M$, and the number of rows of B is $N_c = MN$. A convenient basis for $\mathcal{S}_D$ is the following $MN \times M$ block-diagonal matrix:

$$B_D = \text{diag}\{\mathbf{c}_1(1), \mathbf{c}_1(2), \cdots, \mathbf{c}_1(M)\} \tag{9}$$

where $N \times 1$ vector $\mathbf{c}_1(i)$ is user 1's spreading code during the ith bit interval. It follows that the desired user's signal at the output of the chip-matched filter is given by $B_D\mathbf{u}$,

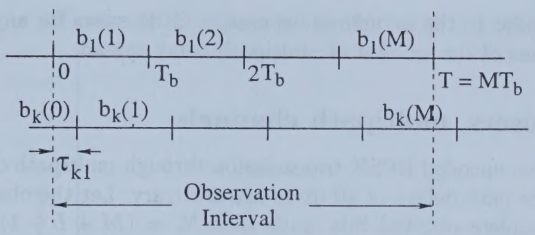


Figure 1: Asynchronous, single path reception.

where $\mathbf{u} = [b_1(1)\ b_1(2)\ \cdots\ b_1(M)]^T$. Therefore, a GLD that is obtained by using this choice for B_D in (8), if one exists, directly yields the decoded bits as its outputs.

Due to asynchronism, $M+1$ independent bits of each interfering user ($b_k(0)$ through $b_k(M)$ in Figure 1) affect the received signal, and therefore the number of dimensions in $\mathcal{S}_I$ due to each interferer is $M+1$. This implies that $N_I = (K-1)(M+1)$. Consider the matched filter output vector $\mathbf{x}_{ki}$ when user k transmits the ith bit and zero everywhere else, and all other users transmit zero everywhere. A natural basis for the interference subspace is the set $\{\mathbf{x}_{ki} : k = 2, 3, \cdots, K;\ i = 0, 1, \cdots, M\}$.

Substituting the above identities in (7), the following condition for the existence of a GLD is obtained:

$$K \leq \frac{MN+1}{M+1}, \tag{10}$$

or, equivalently,

$$K < N \quad \text{and} \quad M \geq \frac{K-1}{N-K}. \tag{11}$$

If (11) is satisfied, then the condition in Proposition 2.1 is satisfied, and a *finite length* GLD can be obtained using (8) for the asynchronous system. (11) shows that as the number of users increases from 1 to $N-1$, the observation interval must be increased from 1 to $N-2$. Also, $M=1$ is sufficient—that is, a one-shot GLD exists—if $K \leq N/2$.

3.2 Synchronous users, multipath channels

Consider synchronous uncoded BPSK transmission through multipath channels. Let $\tau_{k1} = 0$, $k = 1, 2, \cdots, K$, and let the maximum multipath channel spread be L bit periods. Let the observation interval span exactly M complete received bits, such that $N_c = (M+L-1)N$. It can be shown that the dimension of the desired signal space is $N_D = M+2(L-1)$, and the dimension of interference space is $N_I = (K-1)(M+2(L-1))$. Using these in (7), we obtain the following condition for the existence of a GLD:

$$K \leq N\left[1 - \frac{L-1}{M+2(L-1)}\right], \tag{12}$$

or, equivalently,

$$K < N \quad \text{and} \quad M \geq \frac{(L-1)(2K-N)}{N-K}. \tag{13}$$

The above condition (13) shows that for any $K < N$ and any multipath delay spread, there exists an observation interval long enough to allow linear decorrelation. For a fixed delay spread and observation interval, (12) determines the maximum number of users that a GLD can handle. Similar to the asynchronous case, a GLD exists for any $M \geq 1$ whenever $K \leq N/2$, regardless of the amount of multipath delay spread.

3.3 Asynchronous users, multipath channels

Finally, consider asynchronous uncoded BPSK transmission through multipath channels. Because of asynchronism, the path delays of all users are arbitrary. Let the observation interval span exactly M complete received bits, such that $N_c = (M + L - 1)N$. The following limitations of a GLD can be inferred in this case.

$$K \leq N - \frac{NL - 1}{M + 2L - 1}, \tag{14}$$

or, equivalently,

$$K < N \quad \text{and} \quad M \geq \frac{K(2L-1) - N(L-1) - 1}{N - K}. \tag{15}$$

The above results closely parallel those from Section 3.1 and Section 3.2. The case of $M = 1$ is of special interest since it is the usual way in which linear multiuser detectors are implemented. When $M = 1$, (14) implies that the maximum number of users that a GLD can handle is

$$K_{\max} = \begin{cases} \lfloor \frac{N}{2} \rfloor, & L > 1 \\ \lceil \frac{N}{2} \rceil, & L = 1. \end{cases} \tag{16}$$

Therefore, when a one-shot GLD is used, multipath propagation and/or asynchronism decreases the maximum number of users by a factor of two. Also, for any observation interval, as the multipath delay spread increases, the maximum number of users approaches $N/2$. The minimum observation window length M is plotted for different numbers of users K in Figure 2, using $N = 100$ and $L = 2$ in (14). The figure shows that the minimum required window length is equal to one bit period for $K \leq N/2$, and then increases rapidly to a final length of $L(N-2)$ bit periods.

The analyses in Sections 3.1, 3.2 and 3.3 indicate that if a GLD does not exist for $T = (M/N - L + 1)T_c$, then there is no observation interval for which a GLD exists.

4 Effect of Coding or Block Modulation

A drawback of conventional linear multiuser detection is the coding-spreading conflict (see, e.g. [11]). If, for instance, a rate $R < 1$ channel code is used at the transmitter to convert information bits to encoded bits prior to spreading, the maximum number of users for which a conventional linear decorrelator exists is reduced to NR, where N is the spreading gain over information bits. Therefore, any allocation of bandwidth expansion in coding reduces the interference suppression capability of linear decorrelators by the same factor. This loss of linear multiuser detector capability in the presence of coding is due to the fact that conventional linear multiuser detectors do not utilize the knowledge of channel codes. Instead, they treat the encoded signals as the signals to be detected, thereby losing the spreading gain that is used up by coding. The same is true when multi-dimensional modulation is used, that is, $m > 1$ information bits are modulated to one of 2^m symbol waveforms that occupy more than a single dimension

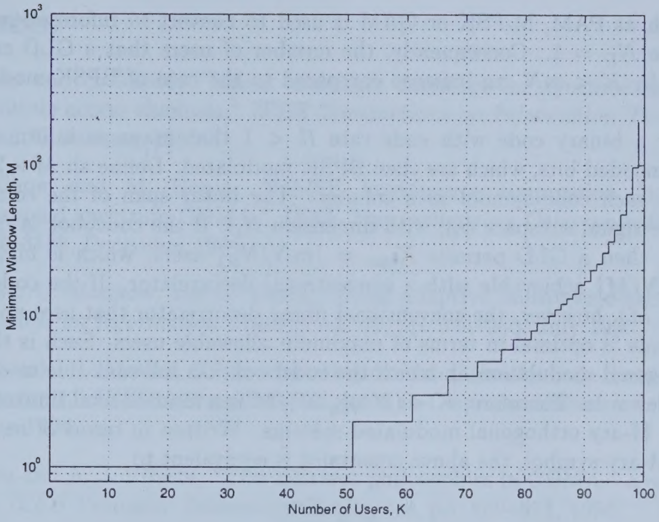


Figure 2: Minimum observation interval versus number of users for $N = 100$ and $L = 2$.

in the signal space. An important example is M-ary orthogonal modulation, in which $\log_2 M$ information bits are block encoded to one of M orthogonal codewords, each of length M. This scheme is used in the uplink of the IS-95 CDMA cellular standard. The use of conventional linear decorrelators for M-ary orthogonal modulated CDMA systems has been suggested in literature (see, e.g., [12]). Since the code rate in the case of M-ary orthogonal modulation is $(\log_2 M)/M$, the number of users that a linear decorrelator can handle is decreased to $(N \log_2 M)/M$, which, for $M = 64$ and $N = 128/3$ (in IS-95; N is over the input bitstream to the M-ary modulator), is merely 4. Evidently, the capability of conventional linear detection is severely limited in this case.

A pertinent question is to see if knowledge of channel codes or modulation scheme can improve the interference suppression capability of linear detectors. This question is addressed here in the framework developed in Section 2. Consider an encoder that maps m information bits to one codeword, which is then multiplied by the spreading code. Let $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{2^m}\}$ be the set of 2^m desired user's codewords as seen at the output of the chip-matched filter. Without loss of generality, we can assume that the observation interval spans the length of exactly one codeword. If the observation interval spans more than one codeword length, the codebook can be redefined to include all possible strings of codewords in the observation interval. The desired signal space $\mathcal{S}_D$ is the space spanned by the above set of code vectors. Its dimension, N_D, determines the interference suppression capability of GLD's, according to the condition in (7). Therefore, knowledge of the space spanned by all codewords can be utilized in designing linear receivers that can potentially overcome some of the limitations of conventional linear multiuser detectors. If all the K users are synchronous, then a GLD exists if, and only if,

$$K < \frac{Nm}{N_D} \tag{17}$$

An interesting observation can be made from (17). If a single dimensional modulation scheme[4] such as PAM, M-PSK or QAM is used to convert m information bits to one symbol, then $N_D = 1$. Consequently, the number of users that a GLD can handle is constrained by $K < mN$, an increase compared to the case of BPSK modulation by a factor of m.

Consider a binary code with code rate $R < 1$ that maps m information bits to $M = m/R$ encoded bits, which are then BPSK modulated. Define an $M \times 2^m$ codebook matrix with each code symbol as a column. The linear span of the columns of this matrix is the signal subspace $\mathcal{S}_D$, with dimension N_D. If the codebook is rank deficient ($N_D < M$), then a GLD permits $K_{\max} = \lfloor mN/N_D \rfloor$ users, which is higher than the $K_{\max} = \lfloor mN/M \rfloor$ achievable with a conventional decorrelator. If the codebook is full rank ($N_D = M$), however, the conventional linear decorrelator that uses no information about the code is optimal in terms of maximum allowable users. Such is the case with M-ary orthogonal modulation, in which the codebook has full rank due to orthogonality of all M codewords. Therefore, $K < (N \log_2 M)/M$ is a fundamental limitation of linear reception in M-ary orthogonal modulated systems. Written in terms of m, the number of bits per M-ary symbol, the above constraint is equivalent to

$$K < \frac{Nm}{2^m}. \tag{18}$$

Therefore, when M-ary orthogonal modulation is used, the interference suppression capability of linear detectors decreases exponentially with increase in block length.

5 Conclusions

This paper showed that though linear multiuser detection is possible in a wide range of scenarios where linear receivers are generally not believed to be applicable, its ability to suppress interference is constrained by the dimensions of spaces spanned by desired and interfering signals. Limitations of linear detection in certain practically important scenarios were studied in a general framework for linear detection derived in this paper. This theory can also be applied to investigate the limitations of linear detectors or to design linear decorrelators in many other situations not addressed in this paper (e.g. presence of coding *and* multipath propagation).

The dimensional limitations of linear multiuser detection, along with their requirement of short codes[5], suggests that nonlinear multiuser detection schemes could be attractive alternatives. We know that there exists at least one nonlinear multiuser detector—the optimal ML detector—that is not affected by the limitations of linear detection schemes discussed in this paper. One attractive solution is an optimal iterative interference canceler [13, 14] that is shown in [14] to approximate minimum probability of error multiuser detection with computational complexity of the order of a conventional parallel interference canceler. Further investigation is required in the analysis of suboptimal nonlinear receivers in various scenarios and performance comparison of linear and nonlinear multiuser detectors.

---

[4] Also called "linear" modulation. Quadrature modulated signals such as QAM also have a single dimension over the complex field.

[5] Neither the current IS-95 CDMA standard nor any of the proposals for wideband CDMA for third generation wireless standards support short codes.

References

[1] R. Lupas and S. Verdú, "Linear multiuser detectors for synchronous code division multiple access channels," *IEEE Transactions on Information Theory*, vol. 35, pp. 123–136, January 1989.

[2] U. Madhow and M. Honig, "MMSE interference suppression for direct sequence spread spectrum CDMA," *IEEE Transactions on Communications*, vol. 42, pp. 3178–3188, December 1994.

[3] M. Honig, U. Madhow, and S. Verdú, "Blind adaptive multiuser detection," *IEEE Transactions on Information Theory*, vol. 41, pp. 944–960, July 1995.

[4] M. Rupf and J. L. Massey, "User-separating demodulation for code-division multiple-access systems," *IEEE Journal Selected Areas in Communications*, vol. 12, pp. 786–795, June 1994.

[5] S. Vembu and A. J. Viterbi, "Two different philosophies in CDMA – a comparison," in *Proc. IEEE Vehicular Technology Conference*, pp. 869–873, 1996.

[6] E. Visotsky and U. Madhow, "Multiuser detection for CDMA systems with nonlinear modulation," in *Proc. International Symposium on Information Theory*, (Ulm, Germany), p. 355, 1997.

[7] V. Ghazi-Moghadam, L. B. Nelson, and M. Kaveh, "Parallel interference cancellation for CDMA systems," in *Proc. 33rd Annual Allerton Conference*, (Urbana-Champaign, IL), 1995.

[8] R. M. Buehrer and B. D. Woerner, "Analysis of multistage interference cancellation for CDMA with phase and timing errors," *IEEE Journal Selected Areas in Communications*, vol. 14, pp. 1522–1535, October 1996.

[9] R. Lupas and S. Verdú, "Near-far resistance of multiuser detectors in asynchronous channels," *IEEE Transactions on Communications*, vol. 38, pp. 496–508, April 1990.

[10] S. Verdú, *Multiuser Detection.* Cambridge University Press, 1998.

[11] V. V. Veeravalli and B. Aazhang, "On the coding-spreading tradeoff in CDMA systems," in *Proc. Conference on Information Sciences and Systems*, (Princeton, NJ), pp. 1136–1141, March 1996.

[12] C. Hegarty and B. Vojcic, "Noncoherent multiuser detection of M-ary orthogonal signals using a decorrelator," in *Proc. Military Communications Conference*, pp. 903–907, 1996.

[13] S. Gollamudi, R. M. Buehrer, S. Nagaraj, and Y.-F. Huang, "Optimal multistage interference cancellation for CDMA systems using the nonlinear MMSE criterion," in *Proc. Asilomar Conference on Signals, Systems and Computers*, (Pacific Grove, CA), November 1998.

[14] S. Gollamudi and Y.-F. Huang, "Iterative nonlinear MMSE multiuser detection," *Submitted to the 1999 International Conference on Acoustics, Speech, and Signal Processing*, March 1999.

Joint Source-Channel Coding of a Gaussian Mixture Source over the Gaussian Broadcast Channel

Zvi Reznic
Dept. of EE-Systems
Tel Aviv University, Israel
and Libit Signal Processing, Israel
zvi@libit.com

Ram Zamir
Dept. of EE-Systems
Tel Aviv University, Israel
zamir@eng.tau.ac.il

Meir Feder
Dept. of EE-Systems
Tel Aviv University, Israel
meir@eng.tau.ac.il

Abstract

Suppose that we want to send a description of a source to two listeners through a Gaussian broadcast channel. The minimum possible distortion D_2, that the bad receiver can obtain is $D_2 = D(C_2)$ where $D(\cdot)$ is the distortion-rate function of the source and C_2 is the capacity of the bad channel. We want to find the smallest achievable distortion D_1, attained by the good receiver, while keeping D_2 optimal. We find this smallest achievable D_1, as well as the entire distortion region, for the case where the source is a Gaussian mixture with highly separated modes. This optimal region exceeds the region attained by separating source and channel coding. Rather, it is attained by an analog-over-digital encoding scheme, in which the "discrete component" of the source is coded and superimposed over the "continuous component" which is sent uncoded.

1 Introduction

The *broadcast channel*, illustrated in Figure 1, is a communication channel in which one sender transmits to two or more receivers.

Suppose we are given a source and a broadcast channel, and we want to convey the source to both receivers simultaneously. Suppose that the source entropy is large to the extent that it cannot be sent losslessly through the channel; there must be some distortion in the receivers' output. The problem of *joint source-channel coding* for the

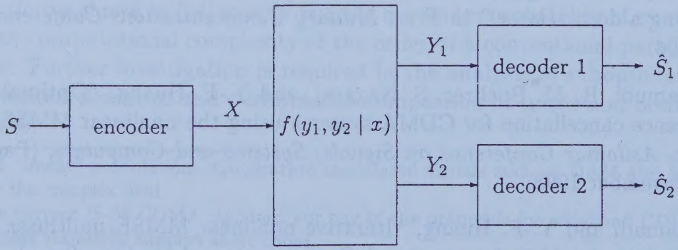


Figure 1: Lossy transmission of a source through a broadcast channel

broadcast channel is to design a communication system to minimize the distortion D_1 at receiver 1, and at the same time minimize the distortion D_2 at receiver 2.

Unfortunately, these two requirements are in conflict. Moreover, for a general source, broadcast channel and distortion measure, it is yet unknown which distortion pairs (D_1, D_2) are simultaneously achievable. We investigate below one example, and find the distortion region.

In our example the channel is a *degraded broadcast channel* in which receiver 1 (the better one) can decode all the information that receiver 2 can, and some additional information. Hence, information decoded at receiver 2 is actually common to both receivers. A rate pair (R_1, R_2) for a degraded broadcast channel is said to be achievable if one can reliably send a common message at rate R_2 to both receivers, and another message at rate R_1 only to the better receiver.

To minimize the distortion at both receivers one might choose an achievable rate pair (R_1, R_2) and then apply a two steps encoding. Step number one is *source coding.* here, we create two messages. The first message, at rate R_2, contains a coarse description of the source. The second message, is a *refinement* message [1, 2]. The refinement message, at rate R_1, is an addendum to the first one, such that both messages together form a fine description of the source. We shall denote by D_c the distortion obtained with the coarse description, and by D_f the distortion obtained with the fine description, that is, $D_f \leq D_c$.

Step number two is *channel coding.* We use a channel code to send the coarse description message to both receivers, and the refinement message to receiver 1 only. Since (R_1, R_2) is achievable, both receivers will correctly receive the coarse description message, and receiver 1 will also correctly receive the refinement message. Hence, (D_f, D_c) is achievable. The two steps encoding is based on *separation*; the source coding and channel coding are done separately.

Unfortunately, unlike for the case of point-to-point communication [4], the distortion pair obtained by the two steps approach is usually suboptimal, e.g. in the simple case of sending a Gaussian source through a Gaussian broadcast channel, with a squared error distortion measure. In this example separation is strictly sub-optimal. In fact, optimality is achieved in this case by analog transmission, i.e., by sending the source *uncoded* [6]. (More insight into this behavior is provided by [7].)

In this paper we present an example where neither separation nor analog transmission are optimal, but rather, a hybrid "digital-over-analog" communication scheme achieves the optimum attainable performance. Our source S, illustrated in Figure 2, is a *Gaussian mixture* source, that is

$$S = aB + N \tag{1}$$

where N is a zero mean Gaussian with variance σ_n^2 and $B \in \{-1, +1\}$ is a binary variable with $Pr(B = 1) = p$, statistically independent of N. Further, we will assume that the two Gaussian "modes" of S are highly separated, i.e., $a \to \infty$.

Our goal is to transmit the source S over a Gaussian broadcast channel and to minimize the mean squared error obtained by the receivers. We apply the digital-over-analog scheme illustrated in Figure 3. This scheme generalizes the concept of *overlaid communication* described in [5]. With this scheme we transmit the Gaussian part of the source analogly, without any coding. At the same time we use source coding and channel coding to digitally transmit the binary part of the source, considering the uncoded transmission as part of the channel's noise. This scheme is simpler than separation based encoding, because part of the source is not encoded. In Theorem 1 we prove that this scheme is

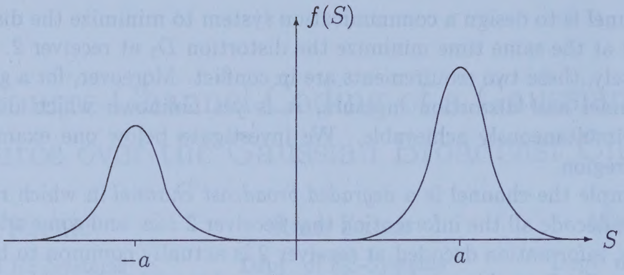


Figure 2: Gaussian mixture source

optimal for point-to-point communication, and in Theorem 2, which is the main result of this paper, we show that the digital-over-analog scheme is also optimal for the broadcast channel, and is strictly better than separation based coding.

2 The Digital-over-Analog Coding Scheme

In this section we present an encoding scheme to transmit a Gaussian mixture source over a Gaussian point-to-point channel. We show that this scheme is optimal with respect to the squared error distortion measure

$$d(s^n, \hat{s}^n) = \frac{1}{n}\sum_{i=1}^{n}(s_i - \hat{s}_i)^2.$$

In section 3 we apply the same scheme to the broadcast channel.

The rate-distortion function of the memoryless Gaussian mixture source S defined in (1), for distortions $D \leq \sigma_n^2$, satisfies [1]

$$R_S(D) = h(S) - \frac{1}{2}\log(2\pi e D) \to H(B) + \frac{1}{2}\log\left(\frac{\sigma_n^2}{D}\right) \qquad \text{as } a \to \infty \tag{2}$$

where H and h denote regular and differential entropy, respectively,

Denote the channel input and output by X and Y respectively, and suppose we have a point-to-point Gaussian channel with average power constraint P, that is

$$\frac{1}{n}\sum_{i=1}^{n} E(X_i^2) \leq P$$

where n is the coding block length,

$$Y = X + Z,$$

and $Z \sim \mathcal{N}(0, \sigma_z^2)$ is statistically independent of X. The capacity of the Gaussian channel [4] is given by:

$$C = \frac{1}{2}\log\left(1 + \frac{P}{\sigma_z^2}\right). \tag{3}$$

---

[1] By the Shannon lower bound $R_S(D) \geq h(S) - \frac{1}{2}\log(2\pi e D)$ for all D, and the lower bound is tight for $D \leq \sigma_n^2$ [3].

By the joint source-channel coding theorem, the minimum achievable distortion D_{min} (the "OPTA") is the solution of $R_S(D_{min}) = C$, implying that

$$D_{min} = \sigma_n^2 \frac{\sigma_z^2}{P + \sigma_z^2} 2^{2H(B)} - \epsilon_a \tag{4}$$

where $\epsilon_a \to 0$ as $a \to \infty$.

The *digital-over-analog* encoder and decoder are illustrated in Figure 3. The source splitter splits each sample S of the source into a binary variable $B' \in (-1, 1)$, with $\Pr(B' = 1) = \Pr(B = 1) = p$, and an independent Gaussian variable $N' \sim \mathcal{N}(0, \sigma_n^2)$, such that $B' + N' = S$. Note that our goal is to transmit S so as long as $B' + N' = S$ it is not important whether B' equals B. The Appendix details the operation of the source splitter. Each sample N' is scaled by a scalar K_1 to produce X_N. A source-channel code i_n encodes the n-block $\mathbf{B}' = (B'_1, \ldots, B'_n)$ to produce the vector $\mathbf{X}_B = i_n(\mathbf{B}')$, which is then added to the vector $\mathbf{X}_N$ to produce the encoder output $\mathbf{X}$. We have

$$Y_j = X_{B_j} + (X_{N_j} + Z_j) \qquad j = 1 \ldots n \tag{5}$$

where for each time instant j, $(X_N + Z) \sim \mathcal{N}(0, P_g + \sigma_z^2)$ where $P_g \triangleq E\left(X_N^2\right)$. Define

$$P_B \triangleq \frac{1}{n} \sum_{j=1}^{n} E X_{B_j}^2 . \tag{6}$$

Equations (5)-(6) specify an equivalent Gaussian channel over which we send B' and whose capacity is given by:

$$C_B = \frac{1}{2} \log \left(1 + \frac{P_B}{\sigma_z^2 + P_g}\right). \tag{7}$$

To transmit B' reliably we choose P_g and P_B such that for some $\Delta > 0$,

$$C_B - \Delta = H(B') = H(B) = -p \log(p) - (1 - p) \log(1 - p). \tag{8}$$

Combining (7),(8) and the power constraint $P_g + P_B = P$ and solving for P_g yields for $\Delta \to 0$:

$$P_g = P_g^* \triangleq \frac{\sigma_z^2 + P}{2^{2H(B)}} - \sigma_z^2 \quad \text{and} \quad K_1 = \sqrt{\frac{P_g^*}{\sigma_n^2}} . \tag{9}$$

The decoder first estimates $\hat{\mathbf{B}}' = g_n(\mathbf{Y})$. Then it generates $\hat{\mathbf{X}}_B$ from $\hat{\mathbf{B}}'$, in the same manner as the encoder generates $\mathbf{X}_B$ from $\mathbf{B}$, i.e., $\hat{\mathbf{X}}_\mathbf{B} = i_n(\hat{\mathbf{B}}')$. The decoder subtracts $\hat{\mathbf{X}}_B$ from the channel output $\mathbf{Y}$ and estimates

$$\hat{\mathbf{N}}' = K_2(\mathbf{Y} - \hat{\mathbf{X}}_B)$$

and

$$\hat{\mathbf{S}} = a\hat{\mathbf{B}}' + \hat{\mathbf{N}}'. \tag{10}$$

The overall mean squared error of the coding scheme is

$$D_{split} \triangleq E\left(\frac{1}{n} \sum_{i=1}^{n} (\hat{S}_i - S_i)^2\right) = E\left(\frac{1}{n} \sum_{i=1}^{n} (a\hat{B}'_i + \hat{N}'_i - aB'_i - N'_i)^2\right). \tag{11}$$

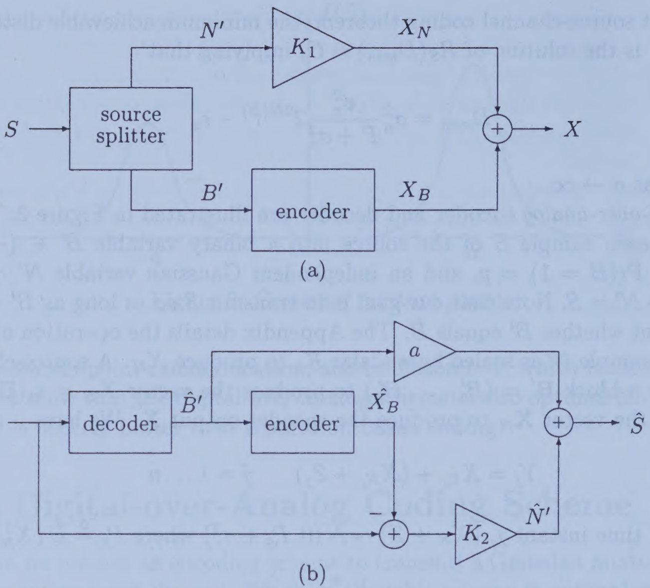


Figure 3: Digital over analog encoding scheme. (a) Encoder. (b) Decoder.

Now, by the source-channel coding theorem, since $H(B) = C_B - \Delta$, there exist i_n and g_n such that

$$\lim_{n\to\infty} Pr(\hat{\mathrm{B}}' \neq \mathrm{B}') = 0. \tag{12}$$

Substituting (12) into (11), we have for $n \to \infty$:

$$D_{split} = \lim_{n\to\infty} E\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{S}_i - S_i)^2 \middle| \hat{\mathrm{B}}' = \mathrm{B}'\right) = E((K_2(K_1 N' + Z) - N')^2) + \epsilon_\Delta \tag{13}$$

where $\epsilon_\Delta \to 0$ as $\Delta \to 0$. Thus the expected distortion is asymptotically the same as the average squared error for a Gaussian source with variance σ_n^2 transmitted uncoded with power P_g^* over a channel with additive white Gaussian noise with power σ_z^2, leading to (with $K_2 = P_g^*/(P_g^* + \sigma_z^2)$)

$$D_{split} = \frac{\sigma_z^2 \sigma_n^2}{P_g^* + \sigma_z^2} + \epsilon_\Delta = D_{min} + \epsilon_a + \epsilon_\Delta. \tag{14}$$

We therefore proved the following theorem.

Theorem 1 *If $C \geq H(B)$ and $a \to \infty$, then the digital-over-analog coding scheme of Figure 3 with K_1, K_2, i_n and g_n as described before, is optimal in the sense that*

$$\lim_{a\to\infty} \lim_{\Delta\to 0} \lim_{n\to\infty} (D_{split} - D_{min}) = 0.$$

The problem of sending a Gaussian mixture source over a Gaussian channel is of practical importance. For example, if one wishes to transmit a song and simultaneously transmit its words in a digital form. In particular, it is practical for $a \to \infty$, since it represents a case where the digital data must be transmitted losslessly. In practice, the digital-over-analog scheme is simpler than the classical separation scheme (source encoder followed by channel encoder) because only a part of the source needs to be encoded and hence it requires less coding effort. This advantage becomes significant when $C \gg H(B)$. Theorem 1 shows that the digital-over-analog scheme is not only simple, it is also optimal.

3 Distortion Region for a Broadcast Channel

We turn to consider the broadcast channel. The goal of the encoder and decoders in Figure 1 is to simultaneously minimize the distortion D_1 at decoder 1 and D_2 at decoder 2. The combination of a certain source, broadcast channel and distortion measure implies a set of distortion pairs that are achievable. We would call this set the *achievable distortion region.* Optimal distortion pairs are those in which decreasing D_1 must imply increasing D_2 and vice versa. Figure 3(a) illustrates a typical shape of the distortion region.

In section 1 we mentioned the fact that analog transmission is optimal for sending Gaussian source over a Gaussian broadcast channel. Hence, the digital-over-analog encoding scheme seems appealing for sending the Gaussian mixture source, as the "Gaussian part" of the information is sent in an analog form (uncoded). To apply the digital-over-analog encoder to the broadcast channel, we should choose the powers P_g and P_B so that receiver 2 will be able to decode B' losslessly. Since receiver 1 is better, it will also decode B' losslessly. As for the Gaussian part, each receiver will obtain a distortion in accordance with the noise level in its channel.

Figure 3(b) shows the distortion region for the Gaussian mixture source S and the Gaussian broadcast channel. The dotted line in Figure 3(b) shows the boundary of the region achievable by separation of source and channel coding as explained in the Introduction. This region is strictly smaller than the optimum region drawn with solid line, which is achieved by our digital-over-analog scheme. We characterize this region in a coding theorem after introducing some formal definitions.

Definition 1 *(D_1, D_2) is an achievable distortion pair for the source S, the distortion measure $d(s, \hat{s})$ and the channel $f(y_1^n, y_2^n \mid x^n)$ if there exist an encoding function $i_n : \mathcal{S}^n \to \mathcal{X}^n$ and two reconstruction functions $g_{1n} : \mathcal{Y}_1^n \to \hat{\mathcal{S}}_1^n$ and $g_{2n} : \mathcal{Y}_2^n \to \hat{\mathcal{S}}_2^n$ such that for $i = 1, 2$*

$$D_i = \lim_{n \to \infty} E\left(d(S^n, \hat{S}_i^n)\right).$$

The achievable distortion region is defined as the set of achievable distortion pairs.

The distortion region is in general convex by a time sharing argument.

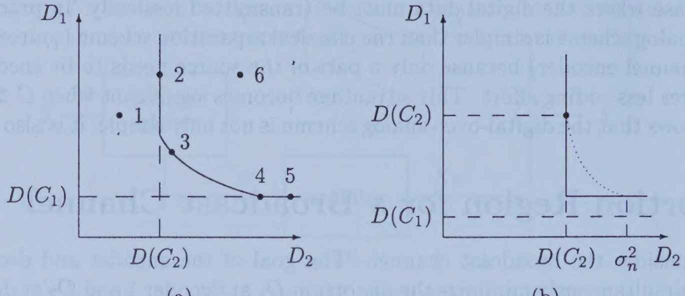
Definition 2 *A* Gaussian *broadcast channel $f(y_1, y_2 \mid x)$ satisfies for $i = 1, 2$:*

$$\frac{1}{n} \sum_{j=1}^{n} E(X_j^2) \leq P,$$

$$Y_i = X + Z_i,$$

$$Z_i \sim \mathcal{N}(0, \sigma_i^2),$$

where Z_1, Z_2 and X are statistically independent of each other.



(a) (b)

Figure 4: Achievable distortion region for the broadcast channel. (a) The distortion region for a general source-channel setting. Point 1 is not achievable. Points 2-6 are achievable. (b) The distortion region of the Gaussian Mixture source transmitted over a Gaussian broadcast channel. The dotted line is the boundry of the (sub-optimal) distortion region achieved by separation. Note that for $a \to \infty$, $D_1 = D(C_1)$ cannot be obtained for any finite D_2.

Theorem 2 *The distortion region for sending the Gaussian mixture source S of (1) in the limit as $a \to \infty$, over a Gaussian broadcast channel with $C_2 > H(B)$, is composed of all pairs (D_1, D_2) satisfying $D_1 \geq D_1^*$ and $D_2 \geq D_2^*$, where*

$$(D_1^*, D_2^*) = \left(\frac{\sigma_1^2 \sigma_n^2}{P_g + \sigma_1^2}, \frac{\sigma_2^2 \sigma_n^2}{P_g + \sigma_2^2} \right) \tag{15}$$

and

$$P_g = \frac{\sigma_2^2 + P}{2^{2H(B)}} - \sigma_2^2.$$

Furthermore, the distortion region is achievable by the digital-over-analog scheme of Figure 3.

Remark: Note that D_2^* in (15) satisfies $R_S(D_2^*) = C_2$.

To prove the converse part of Theorem 2 we present two lemmas. The first lemma provides a lower bound on D_1 for any a. An extension of this lemma for more general sources and distortion measures is given in [3].

Lemma 1 *Suppose we are given some joint source-channel coding scheme. Let P_{e2} be the minimum average probability of error in estimating the vector $\mathbf{B}$ from the reconstruction vector $\hat{\mathbf{S}}_2$ of the second (bad) receiver, and let*

$$\delta = H_b(P_{e2}) \tag{16}$$

where H_b denotes the binary entropy. Then, the rate distortion function of the source at distortion D_1 (the distortion level of the first receiver), satisfies:

$$R_S(D_1) \leq H(B) + \frac{1}{2} \log \left(1 + \frac{P_\delta}{\sigma_1^2} \right) \tag{17}$$

where

$$P_\delta \triangleq \frac{\sigma_2^2 + P}{2^{2(H(B)-\delta)}} - \sigma_2^2.$$

Note that since $R_S(D)$ is monotonically non-increasing, (17) provides a lower bound for the distortion D_1 of the good receiver, depending on how well the bad receiver detects the source mode B'.

Proof: By the Markov chain $B \leftrightarrow S \leftrightarrow X \leftrightarrow Y_2 \leftrightarrow \hat{S}_2$ and by Fano's inequality we have

$$H(\mathbf{B}|\mathbf{S}) \leq H(\mathbf{B}|\mathbf{X}) \leq H(\mathbf{B}|\mathbf{Y_2}) \leq H(\mathbf{B}|\hat{\mathbf{S}}_2) \leq n\delta. \tag{18}$$

By the chain rule for mutual information we have for $i = 1, 2$:

$$\begin{aligned} I(\mathbf{X};\mathbf{Y_i}) &= I(\mathbf{X};\mathbf{B}) + I(\mathbf{X};\mathbf{Y_i}|\mathbf{B}) - I(\mathbf{X};\mathbf{B}|\mathbf{Y_i}) \\ &= I(\mathbf{X};\mathbf{B}) + h(\mathbf{Y_i}|\mathbf{B}) - h(\mathbf{Z_i}) - I(\mathbf{X};\mathbf{B}|\mathbf{Y_i}). \end{aligned} \tag{19}$$

Using (19) with $i = 2$ we upper bound $h(\mathbf{Y_2}|\mathbf{B})$ as follows:

$$\begin{aligned} nC_2 &\geq I(\mathbf{X};\mathbf{Y_2}) = I(\mathbf{X};\mathbf{B}) + h(\mathbf{Y_2}|\mathbf{B}) - h(\mathbf{Z_2}) - I(\mathbf{X};\mathbf{B}|\mathbf{Y_2}) \\ &= H(\mathbf{B}) - H(\mathbf{B}|\mathbf{X}) + h(\mathbf{Y_2}|\mathbf{B}) - h(\mathbf{Z_2}) - H(\mathbf{B}|\mathbf{Y_2}) + H(\mathbf{B}|\mathbf{X},\mathbf{Y_2}) \\ &\geq n\left(H(B) + \frac{1}{n}h(\mathbf{Y_2}|\mathbf{B}) - h(Z_2) - \delta\right) \end{aligned} \tag{20}$$

where we used (18) to upper bound $\frac{1}{n}H(\mathbf{B}|\mathbf{Y_2})$ by δ. Using (19) with $i = 1$ we lower bound $h(\mathbf{Y_1}|\mathbf{B})$ as follows:

$$\begin{aligned} nR_s(D_1) &\leq I(\mathbf{S};\hat{\mathbf{S}}_1) \leq I(\mathbf{X};\mathbf{Y_1}) \qquad (21) \\ &= I(\mathbf{X};\mathbf{B}) + h(\mathbf{Y_1}|\mathbf{B}) - h(\mathbf{Z_1}) - I(\mathbf{X};\mathbf{B}|\mathbf{Y_1}) \\ &\leq I(\mathbf{X};\mathbf{B}) + h(\mathbf{Y_1}|\mathbf{B}) - h(\mathbf{Z_1}) \\ &\leq n\left(H(B) + \frac{1}{n}h(\mathbf{Y_1}|\mathbf{B}) - h(Z_1)\right). \end{aligned}$$

where we used $I(X;B) \leq H(B)$ and the nonnegativity of the mutual information. Finally by the entropy power inequality:

$$2^{\frac{2}{n}h(\mathbf{Y_2}|\mathbf{B})} \geq 2^{\frac{2}{n}h(\mathbf{Y_1}|\mathbf{B})} + 2^{\log(2\pi e(\sigma_2^2 - \sigma_1^2))}. \tag{22}$$

Combining (20), (21) and (22) and using $C_2 = \frac{1}{2}\log\left(1 + \frac{P}{\sigma_2^2}\right)$ yields the desired result. □

Lemma 2 *Let* P_{e2} *be defined as in Lemma 1 above, and let* D_2 *be the mean squared error of the bad receiver. For any coding scheme,* $P_{e2} \leq \epsilon(a, D_2)$*, where* $\epsilon(a, D_2) \to 0$ *as* $a \to \infty$ *for any* $D_2 < \infty$.

The Lemma is proved in [3].

Proof of Theorem 2

(Achievability:) Suppose we use the digital-over-analog scheme, where we choose the power P_g of the Gaussian part such that the bad receiver (and obviously the good receiver) can decode B' losslessly, and we choose the gain K_2 of each receiver according to its own noise level. Then from (9)

$$P_g = \frac{\sigma_2^2 + P}{2^{2H(B)}} - \sigma_2^2,$$

and from (14) as $n \to \infty$

$$(D_1, D_2) = \left(\frac{\sigma_1^2 \sigma_n^2}{P_g + \sigma_1^2}, \frac{\sigma_2^2 \sigma_n^2}{P_g + \sigma_2^2} \right) = (D_1^*, D_2^*).$$

(Converse:) By Lemma 2, for any fixed $D_2 < \infty$ we have $P_{e2} \to 0$ as $a \to \infty$, hence $\delta \to 0$ in (16). Substituting in (17) and combining with (2) yields $D_1 \geq D_1^*$. As for D_2, by Shannon's joint source-channel coding theorem $D_2 \geq D_2^*$ for any D_1, since $R_S(D_2^*) = C_2$. This proves the converse. $\square$

4 Conclusions

The digital-over-analog method splits the source into two parts, one is coded and the other is transmitted uncoded. (The scheme can be extended to the case where B and N are dependent, using a *lookahead* method [8].) The digital-over-analog scheme is simpler than traditional methods which are based on separating the source coding from the channel coding. For the Gaussian mixture source and Gaussian channel, this method was proved to be optimal both for point-to-point communication and for broadcast channel.

The distortion region for the broadcast channel is yet unknown. Here we found it in one special case.

Appendix: The source splitter

The source splitter needs to split the source S to its Gaussian component and binary component. Practically, if $a >> \sigma_n^2$, we can set $B' = \text{sign}(S)$ and $N' = S - aB'$. However, with this splitting method the p.d.f. of N' given B' is limited on one side, and therefore N' is not Gaussian. So, for mathematical accuracy, the source splitter chooses B' as follow. For each input $S = s$ the source splitter flips an unfair coin with

$$\Pr(head) = \frac{K_0(s)}{K_0(s) + K_1(s)}$$

where

$$K_0(s) \triangleq p \frac{1}{\sigma_n} e^{-(s-a)^2/2\sigma_n^2},$$

and

$$K_1(s) \triangleq (1-p) \frac{1}{\sigma_n} e^{-(s+a)^2/2\sigma_n^2}.$$

If the coin lies on head then the source splitter sets $B' = 1$, else it sets $B' = -1$. N' is calculated by $N' = S - aB'$. It is easy to prove that $\Pr(B' = 1) = p$, $N' \sim \mathcal{N}(0, \sigma_n^2)$ and B' and N' are statistically independent.

References

[1] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory,* vol. IT-37, no. 2, pp. 269-274, Mar. 1991.

[2] B. Rimoldy, "Successive refinement of information: characterization of the achievable rates," *IEEE Trans. Inform. Theory,* vol. IT-40, no. 1, pp. 253-259, Jan. 1994.

[3] Z. Reznic R. Zamir and M. Feder, "Joint source-channel coding of a Gaussian mixture source over the Gaussian broadcast channel," to be submitted to *IEEE. Trans. Inform. Theory*

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[5] S. Shamai, S. Verdú and R. Zamir, "Systematic lossy source-channel coding", *IEEE Trans. Inform. Theory,* vol. IT-44, no. 2, pp. 564-579, March 1998.

[6] T.J. Goblick, "Theoretical limitations on the transmission of data from analog sources", *IEEE Trans. Inform. Theory,* vol. IT-11, pp. 558-567, Oct. 1965.

[7] J. Ziv, "The behavior of analog communication systems", *IEEE Trans. Inform. Theory,* vol. IT-16, pp. 587-594, Sept. 1970.

[8] S. Shamai and S. Verdú, "Capacity of channels with uncoded side information," *Europ. Trans. Telecommun.,* vol. 6, no. 5, pp 587-600, Sept.-Oct. 1995.

Separation of Random Number Generation and Resolvability *

K. Visweswariah, S. Kulkarni, S. Verdú

Department of Electrical Engineering
Princeton University, Princeton, NJ 08544

Abstract

We consider the problem of when a given source can approximate the output due to any input to a given channel. We provide necessary and sufficient conditions for a general source and channel. For the special case of a full rank discrete memoryless channel we give a stronger necessary condition than for a general channel.

1 Introduction

The classical separation theorem essentially states that source and channel coding can be done separately without losing optimality. In [1] a Source-Channel separation theorem was shown that is more general than the classical separation theorem. Here we investigate an analogous separation theorem in the case of the resolvability and the random number generation. Random number generation involves finding a deterministic transformation to generate a sequence of equiprobable bits from a given source of randomness. Resolvability is a property of a channel which gives the amount of randomness required to simulate, at the output of the channel, any distribution that can be achieved by a random input to the channel. We consider the two problems together, i.e. when can a given source of randomness be used to simulate the output of a given channel due to any input to arbitrary accuracy. In Section 2 we formalize the problem we consider. Sections 3 and 4 give sufficient and necessary conditions for a source to approximate the output of a given channel due to any input. Section 5 deals with the special case when the channel is a discrete memoryless channel of full rank. For this special case we will be able to show a necessary condition which is stronger than the necessary conditions in Section 4 for a general channel.

2 Preliminaries and Problem Formulation

In this section we will give some basic definitions, give a precise statement of the problem we are looking at and give results which are already known for the problem of separation of resolvability and random number generation.

---

*This work was partially supported by the National Science Foundation under Grants NYI Award IRI-9457645 and NCR 9523805

This definition is as in [3] and we repeat it for the sake of completeness.

Definition 1 *A channel* $\mathbf{W}$ *with input alphabet* A *and output alphabet* B *is a sequence of conditional distributions*

$$\mathbf{W} = \{W^n(y^n|x^n); (x^n, y^n) \in A^n \times B^n\}_{n=1}^{\infty}.$$

Definition 2 *A source of randomness* $\mathbf{X}$ *is a sequence of finite dimensional distributions* $\{P_{X^n}\}_{n=1}^{\infty}$ *with* X^n *taking values in* A^n.

Note that there are no consistency requirements on the sequence of finite dimensional distributions, this is the difference between Definition 2 and the standard definition of a random process. Throughout we assume that the source and channel alphabets are finite.

We now give some notation that use throughout this paper. Let the output distribution, when the input is distributed according to Q^n be denoted by Q^nW^n, thus

$$Q^nW^n(y^n) = \sum_{x^n \in A^n} W^n(y^n|x^n)\, Q^n(x^n).$$

Also let

$$i_{X^nW^n}(a^n, b^n) \triangleq \log \frac{W^n(b^n|a^n)}{P_{Y^n}(b^n)}$$

(where P_{Y^n} is the output distribution when P_{X^n} is input to the channel W^n) and

$$h_{Z^n}(z^n) \triangleq \log \frac{1}{P_{Z^n}(z^n)}.$$

We will use l_1 distance to measure the difference between two distributions on the same alphabet. We will denote the distance between P and Q by $d(P, Q)$. We note that $d(P, Q) = 2\sup_{E \subseteq A} |P(E) - Q(E)|$ where A is the alphabet on which the two distributions are defined.

We now define precisely what we mean by a source $\mathbf{Z}$ being an approximating source for a channel $\mathbf{W}$.

Definition 3 *For any* $\epsilon > 0$, *the source* $\mathbf{Z}$ *with alphabet* F *is called an* ϵ*-approximating source for the channel* $\mathbf{W}$ *if for any arbitrary input source* $\tilde{\mathbf{X}}$, *there exists a sequence of deterministic mappings* $\{\phi_n : F^n \longmapsto A^n\}$ *such that for sufficiently large* n,

$$d\left(Y^n, \tilde{Y}^n\right) < \epsilon$$

where Y^n *and* $\tilde{Y}^n$ *are the outputs of the channel due to* $\phi(Z^n)$ *and* $\tilde{X}^n$ *respectively.*

Using this definition of an ϵ-approximating source we can now define $\mathbf{Z}$ to be an *approximating source* for $\mathbf{W}$ if it is ϵ-approximating for $\mathbf{W}$ for all $\epsilon > 0$.

The problem of resolvability of a channel was considered in [3]. The resolvability of a channel is the minimum number of random bits required per input sample to approximate arbitrarily well the output of the channel due to any input process (See [3] for a formal definition of resolvability). It is shown in [3] that the resolvability of a channel is $\sup_{\mathbf{X}} \overline{I}(\mathbf{X}, \mathbf{Y})$. The problem we look at here is that of finding necessary and sufficient conditions for a given source $\mathbf{Z}$ to be an approximating source for a channel $\mathbf{W}$. To use a given source to approximate the output of a channel due to another source we could use the source to generate random bits at the best possible rate and then use a deterministic transformation of the random bits at the input of the channel. The problem of random bit generation was considered in [2] where fundamental limits on the rate at which random bits can be generated from a given source were given. Using the results of [3] and [2] and a two step process outlined above we can easily show the following sufficient condition:

Theorem 1 *If $\underline{H}(\mathbf{Z}) > S$ then $\mathbf{Z}$ is an approximating source for a channel $\mathbf{W}$, where S is the resolvability of the channel $\mathbf{W}$.*

The necessary condition below can also be derived using the results in [3]. A source $\mathbf{Z}$ can be approximated using $\overline{H}(\mathbf{Z})$ random bits per sample (Theorem 3, [3]). This also means any process derived from $\mathbf{Z}$ by a deterministic transformation can be approximated using $\overline{H}(\mathbf{Z})$ random bits per sample. Thus if a source $\mathbf{Z}$ is an approximating source for a channel then we can approximate any output of the channel with $\overline{H}(\mathbf{Z})$ random bits per sample and so the resolvability S of the channel must be smaller than $\overline{H}(\mathbf{Z})$. Thus we have the following theorem:

Theorem 2 *If $\overline{H}(\mathbf{Z}) < S$ then $\mathbf{Z}$ is not an approximating source for a channel $\mathbf{W}$.*

In the next two sections we will give stronger versions of the above two theorems.

3 A Sufficient Condition

In this section we give a sufficient condition for a source $\mathbf{Z}$ to be an approximating source for a channel $\mathbf{W}$ which implies the sufficient condition given by Theorem 1. The sufficient condition is analogous to sufficient condition derived in [1] for the Source-Channel separation problem. We will define first the notion of a source strictly dominating a channel analogous to the definition in [1].

Definition 4 *A source $\mathbf{Z}$ is said to strictly dominate a channel $\mathbf{W}$ if for every channel input process $\mathbf{X}$ there exists a $\delta > 0$ such that*

$$\lim_{n\to\infty} \inf_{c_n \in R} \left\{ P\left[\frac{1}{n} h_{Z^n}(Z^n) \le c_n + \delta\right] + P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \ge c_n\right] \right\} = 0.$$

We note that $\underline{H}(\mathbf{Z}) > S$ implies that $\mathbf{Z}$ dominates $\mathbf{W}$ since if we take $c_n = (\underline{H}(\mathbf{Z}) + S)/2$ the required condition is satisfied. The converse statement however, is not true.

Theorem 3 *If a source $\mathbf{Z}$ strictly dominates a channel $\mathbf{W}$ then $\mathbf{Z}$ is an approximating source for the channel $\mathbf{W}$.*

Proof The source $\mathbf{Z}$ strictly dominates the channel $\mathbf{W}$ implies that for each input process $\mathbf{X}$ there exists a sequence $\{c_n\}_{n=1}^{\infty}$ and a $\delta > 0$ such that

$$P\left[\frac{1}{n} h_{Z^n}(Z^n) \le c_n + \delta\right] \le \tau_n \tag{1}$$

and

$$P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \geq c_n\right] \leq \tau_n \tag{2}$$

where $\tau_n \to 0$ as $n \to \infty$. Equation (1) implies that we can approximate any distribution with type less than $\exp n\left(c_n + \frac{2}{3}\delta\right)$ using the distribution Z^n. To show this we use the procedure in the Aggregation Lemma [2].

Define $S^{(n)} = \left\{z^n \in F^n : P_{Z^n}(z^n) \leq \exp^{-n(c_n+\delta)}\right\}$. Consider any $M-$type distribution P_M on $\{1, 2, \ldots, M\}$. We will place elements of $S^{(n)}$ in bin $B_n(i)$ until we have

$$P_{Y^n}(B_n(i)) > P_M(i) - \exp^{-n(c_n+\delta)}.$$

We stop either when we complete this process for all $i = 1, 2, \ldots, M$ or when we run out of sequences in the set $S^{(n)}$. All remaining sequences in F^n are placed in $B_n(1)$. At the end of this process we have

$$\sum_{i=1}^{M} |P_{Z^n}(B_n(i)) - P_M(i)| \leq \max\left(2M \exp^{-n(c_n+\delta)}, 2\tau_n + M \exp^{-n(c_n+\delta)}\right)$$

For any $M \leq \exp n\left(c_n + \frac{2}{3}\delta\right)$ the right hand side of the last equation goes to zero, since $\tau_n \to 0$.

We will now state and use a lemma the proof of which readily follows from the argument used in the proof of Theorem 4, [3].

Lemma 1 *If*

$$P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \geq c_n\right] \leq \tau_n$$

where $\tau_n \to 0$ as $n \to \infty$ then for any $\gamma > 0$ there exists a process $\tilde{\mathbf{X}}$ such that

$$\lim_{n\to\infty} d\left(Y^n, \tilde{Y}^n\right) = 0$$

and $\tilde{X}^n$ is an $M-$type distribution with $M \leq \exp n\,(c_n + \gamma)$

Equation (2) and Lemma 1 imply that there exist a process $\tilde{X}^n$ with type smaller than $\exp n\left(c_n + \frac{2}{3}\delta\right)$ which approximates the output due to $\mathbf{X}$. We can approximate arbitrarily closely any distribution with type less than or equal to $\exp n\left(c_n + \frac{2}{3}\delta\right)$ (and hence $\tilde{X}^n$) using Z^n. This along with the fact that $d(PW, QW) \leq d(P, Q)$ for any channel W and distributions P, Q imply that $\mathbf{Z}$ can be used to approximate the output of the channel $\mathbf{W}$ when the input process is $\mathbf{X}$. Since we can find a sequence $\{c_n\}_{n=1}^{\infty}$ and a $\delta > 0$ which satisfy equations (1) and (2) for any input process $\mathbf{X}$ the source $\mathbf{Z}$ can approximate the output due to any process $\mathbf{X}$.

□

4 Necessary Condition

In this section we will give necessary conditions for a source $\mathbf{Z}$ to be an approximating source for a channel $\mathbf{W}$. Again we would like to have a result completely analogous to the necessary condition in [1] for the Source-Channel separation problem. To state that necessary condition we will need the notion of a source $\mathbf{Z}$ dominating as opposed to strictly dominating a channel $\mathbf{W}$.

Definition 5 *A source* $\mathbf{Z}$ *dominates the channel* $\mathbf{W}$ *if for every process* $\mathbf{X}$*, for every* $\delta > 0$ *and for every sequence of non-negative numbers* $\{c_n\}_{n=1}^{\infty}$

$$\lim_{n\to\infty} P\left[\frac{1}{n}h_{Z^n}(Z^n) \le c_n - \delta\right] P\left[\frac{1}{n}i_{X^nW^n}(X^n;Y^n) \ge c_n\right] = 0.$$

It can be verified that if a source strictly dominates a channel then it dominates the channel. Also note that $\overline{H}(\mathbf{Z}) < S$ implies that the source does not dominate the channel. The final result we would like to have would be: $\mathbf{Z}$ is an approximating source for a channel $\mathbf{W}$ implies that the source dominates the channel which would be a stronger statement than Theorem 2. We have not been able to show this statement but we give two necessary conditions, one of which gives a stronger statement than Theorem 2 and the other neither implies nor is implied by Theorem 2.

Theorem 4 *Suppose that for some process* $\mathbf{X}$*, for some* $\alpha, \delta > 0$ *and for some sequence of non-negative numbers* $\{c_n\}_{n=1}^{\infty}$

$$P\left[\frac{1}{n}h_{Z^n}(Z^n) \le c_n - \delta\right] > \beta_n \tag{3}$$

and

$$P\left[\frac{1}{n}i_{X^nW^n}(X^n;Y^n) \ge c_n\right] > \alpha \tag{4}$$

for all $n \in I$ *where* I *is an infinite set of integers and* $\beta_n \to 1$ *as* $n \to \infty$*. Then the source* $\mathbf{Z}$ *is not an approximating source for the channel* $\mathbf{W}$*.*

Proof In this proof we will concentrate on $n \in I$. We will prove this theorem by contradiction. Assume Equations (3),(4) are satisfied and that $\mathbf{Z}$ is an approximating source for the channel $\mathbf{W}$.

We note that if $V^n = \phi^n(Z^n)$ and if Z^n satisfies Equation (3) then V^n also satisfies the same property. This implies that for any deterministic transformation ϕ^n, $\phi^n(Z^n)$ can be approximated arbitrarily closely for sufficiently large $n \in I$ by an $M = \exp n\left(c_n - \frac{\delta}{2}\right)$ type distribution. To see this consider V^n satisfying

$$P\left[\frac{1}{n}h_{V^n}(V^n) \le c_n - \delta\right] > \beta_n$$

where $\beta_n \to 1$ as $n \to \infty$. Let

$$B^n = \left\{v^n : \frac{1}{n}h_{V^n}(V^n) \le c_n - \delta\right\}.$$

Clearly $|B^n| \le \exp n(c_n - \delta)$. We can approximate the distribution of V^n to within $2\frac{|B^n|}{M} + (1 - P_{V^n}(B^n))$ with an M type distribution. Thus the distribution P_{V^n} can be approximated to within $\exp -\left(n\frac{\delta}{2}\right) + (1 - \beta_n)$ by an $M = \exp n\left(c_n - \frac{\delta}{2}\right)$ type distribution.

Now using a proof similar to the proofs of Lemmas 7 and 8 in [3] Equation (4) implies that for sufficiently small $\epsilon > 0$ we can find N distributions $\{Q_i\}_{i=1}^{N}$ with

$$\frac{1}{n}\log\log N > c_n - \frac{\delta}{10}$$

and

$$\min_{i \neq j} d\left(Q_i^n W^n, Q_j^n W^n\right) \geq 2\epsilon$$

for all sufficiently large $n \in I$.

Input distributions which are ϵ-apart result can result in output distributions which are at most ϵ-apart. We note that the Definition of an ϵ-approximating source in Definition 3 is the same as its uniform version where the "sufficiently large n" for which the statement holds is independent of $\tilde{\mathbf{X}}$. This follows from a proof similar to that in Lemma 6 of [3]. Thus if Z^n is an approximating source for the channel then we must have for sufficiently large $n \in I$ at least

$$\exp\left(\exp n\left(c_n - \frac{\delta}{10}\right)\right)$$

distributions derived by deterministic transformations from Z^n, no two of which are less than 2ϵ-apart.

We have already seen that for any $\epsilon > 0$, any distribution derived by a deterministic transformation from Z^n can be approximated to within ϵ by an $M = \exp n\left(c_n - \frac{\delta}{2}\right)$ type distribution. The number of these M type distributions is upper bounded by

$$|A^n|^{\exp n\left(c_n - \frac{\delta}{2}\right)}.$$

Which for sufficiently large n is less than

$$\exp\left(\exp n\left(c_n - \frac{\delta}{10}\right)\right)$$

Thus the number of distributions derived by deterministic transformations from Z^n is not enough to approximate outputs of the channel to all possible inputs and so we have a contradiction.

□

We note that the conditions in Equations (3), (4) imply the negation of the condition that the source $\mathbf{Z}$ dominates the channel $\mathbf{W}$ and thus the necessary condition we have is weaker than the necessary condition analogous to the separation for Source-Channel separation problem in [1]. The result is stronger than that in Theorem 2.

We now give a stronger definition of an approximating source and give another necessary condition for a source to be a strong approximating source.

Definition 6 *For any $\epsilon > 0$, the source $\mathbf{Z}$ with alphabet F is called a strong $\epsilon-$ approximating source for the channel $\mathbf{W}$ if for any arbitrary input source $\tilde{\mathbf{X}}$, there exists a sequence of deterministic mappings $\{\phi_n : F^n \longmapsto A^n\}$ such that $P_{\phi_n(Z^n)} \ll P_{\tilde{X}^n}$ and such that for sufficiently large n,*

$$d\left(Y^n, \tilde{Y}^n\right) < \epsilon$$

where Y^n and $\tilde{Y}^n$ are the outputs of the channel due to $\phi(Z^n)$ and $\tilde{X}^n$ respectively.

We call $\mathbf{Z}$ a *strong approximating source* for $\mathbf{W}$ if it is a strong $\epsilon-$approximating source for $\mathbf{W}$ for all $\epsilon > 0$.

Note that the only difference between Definitions 3 and 6 is that the latter places an extra constraint that the deterministic transformation can only map to those sequences which have non-zero probability under the source whose output we are trying to approximate.

The next theorem states precisely a necessary condition for the source to be a strong approximating source for a channel. Again the statement is in the contra-positive form.

Theorem 5 *Suppose that for some process* $\mathbf{X}$*, for some* $\alpha, \delta > 0$ *and for some sequence of non-negative numbers* $\{c_n\}_{n=1}^{\infty}$

$$P\left[\frac{1}{n} h_{Z^n}(Z^n) \le c_n - \delta\right] > \alpha \tag{5}$$

and

$$P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \ge c_n\right] > \beta_n \tag{6}$$

for all $n \in I$ *where* I *is an infinite set of integers and* $\beta_n \to 1$ *as* $n \to \infty$*. Then the source* $\mathbf{Z}$ *is not a strong approximating source for the channel* $\mathbf{W}$.

Proof Equation (6) along with Feinstein's Lemma [6] implies that there exists a code of length n with $\exp n\left(c_n - \frac{\delta}{10}\right)$ codewords and with maximal probability of error less than ϵ_n for $n \in I$ where $\epsilon_n \to 0$ as $n \to \infty$. Consider a good code with $M = \exp n\left(c_n - \frac{\delta}{10}\right)$ codewords. Let the codewords be $\{b_i\}_{i=1}^{M}$ and their corresponding decoding sets be $\{D_i\}_{i=1}^{M}$. Consider now the process $\tilde{X}^n$ which places mass $\frac{1}{M}$ on each of the M codewords, $\{b_i\}_{i=1}^{M}$. We will show that the output due to this process cannot be approximated well with our source.

Consider the set $S^n = \left\{z^n : \frac{1}{n} h_{Z^n}(Z^n) \le c_n - \delta\right\}$. By Equation (5) $P_{Z^n}(S^n) > \alpha$ for all $n \in I$. Consider the N codewords to which the sequences in S^n get mapped . Assume the codewords are numbered so that these codewords are $\{b_i\}_{i=1}^{N}$. Since $|S^n| \le \exp n\,(c_n - \delta)$, we have $N \le \exp n\,(c_n - \delta)$.

Let

$$B = \bigcup_{i=1}^{N} D_i$$

and $\tilde{Y}^n$ be the output due to $\tilde{X}^n$.

$$\begin{aligned} P_{\tilde{Y}^n}(B) &= \frac{1}{M}\sum_{i=1}^{N} P(B|b_i) + \frac{1}{M}\sum_{i=N+1}^{M} P(B|b_i) \\ &\le \frac{N}{M} + \frac{\epsilon_n}{M}(M - N) \\ &\le 2\epsilon_n \end{aligned}$$

where the last inequality holds for sufficiently large $n \in I$. Now let $\mathbf{X}$ be any process derived by a deterministic transformation from $\mathbf{Z}$ and let $\mathbf{Y}$ be the output process due to the process with $\mathbf{X}$ as the input to the channel.

$$P_{Y^n}(B) = \sum^{M} P(B|b_i)\, P_{X^n}(b_i)$$

$$
\begin{aligned}
&\geq \sum_{i=1}^{N} P(D_i|b_i)\, P_{X^n}(b_i) \\
&\geq (1-\epsilon_n)\, P_{Z^n}(S^n) \\
&\geq (1-\epsilon_n)\, \alpha.
\end{aligned}
$$

Thus we have

$$P_{\tilde{Y}^n}(B) - P_{Y^n}(B) \geq (1-\epsilon_n)\,\alpha - 2\epsilon_n$$

for all sufficiently large $n \in I$. We note that the R.H.S of the equation above does not go to zero as n increases and so $\mathbf{Z}$ cannot be an approximating source in the strong sense for the channel $\mathbf{W}$.

□

5 A Special Case

We will now look at the special case where the channel is a full rank, discrete, memoryless channel (FRDMC) . A channel $\mathbf{W}$ is of full rank if the transition vectors $\{W(.|a)\}_{a\in A}$ are linearly independent. For the FRDMC we will be able to show a result analogous to [1].

Theorem 6 *For a FRDMC if the source* $\mathbf{Z}$ *is an approximating source for the channel* $\mathbf{W}$ *then the source* $\mathbf{Z}$ *dominates the channel* $\mathbf{W}$.

Proof We will show that if the source $\mathbf{Z}$ does not dominate the FRDMC $\mathbf{W}$ then $\mathbf{Z}$ is not an approximating source for the channel $\mathbf{W}$. If a source $\mathbf{Z}$ does not dominate $\mathbf{W}$ then there is a process $\overline{\mathbf{X}}$ such that for some $\alpha, \delta > 0$ and for some sequence of non-negative numbers $\{c_n\}_{n=1}^{\infty}$

$$P\left[\frac{1}{n}h_{Z^n}(Z^n) \leq c_n - \delta\right] > \alpha \tag{7}$$

and

$$P\left[\frac{1}{n}i_{\overline{X}^n W^n}\left(\overline{X}^n;\overline{Y}^n\right) \geq c_n\right] > \alpha \tag{8}$$

for all $n \in I$ where I is an infinite set of integers. From Corollary 2 in [4] we have that for a FRDMC $S = C$. But for a DMC, $C = \sup_X I(X;Y)$. Since $S = \sup_{\mathbf{X}} \overline{I}(\mathbf{X};\mathbf{Y})$ we have for a FRDMC $\sup_{\mathbf{X}} \overline{I}(\mathbf{X};\mathbf{Y}) = \sup_X I(X;Y)$. Thus for a FRDMC if there is a process $\overline{\mathbf{X}}$ such that Equations (7) and (8) are satisfied then there exists an i.i.d process $\mathbf{X}$ such that

$$c_n \leq I(X;Y) + \frac{\delta}{10} \tag{9}$$

for sufficiently large $n \in I$. We will show that the source $\mathbf{Z}$ cannot approximate the output due to process $\mathbf{X}$.

Define the set of sequences in A^n that are not γ typical ($\gamma > 0$) by

$$D_{X^n}(\gamma) = \left\{x^n : \left|\frac{1}{n}N(a|x^n) - P_X(a)\right| > \gamma \text{ for some } a \in A\right\}$$

where $N(a|x^n)$ denotes the number of times that a occurs in the sequence x^n.

Also define the set of sequences jointly typical with x^n by

$$T_W^n(x^n,\gamma) = \left\{y^n : \left|\frac{1}{n}N(a,b|x^n,y^n) - \frac{1}{n}N(a|x^n)W(b|a)\right| \leq \gamma \text{ for all } (a,b) \in A \times B\right\}$$

where $N(a,b|x^n,y^n)$ denotes the number of times that (a,b) occurs in (x^n,y^n).

Let W denote the matrix whose columns are the the transition vectors $\{W(.|a)\}_{a\in A}$. Let $\mathbf{r}$ be an $|A|$ dimensional vector and let $\mathbf{s} = W\mathbf{r}$. Since W is of full rank if $|r_i| > \gamma$ for some $i \in \{1,2,\ldots,|A|\}$ then $|s_j| > k\gamma$ for some $j \in \{1,2,\ldots,|B|\}$ and some $k > 0$ independent of $\mathbf{r}$. Thus we have that if $x^n \in D_X^n(\gamma)$ then $W\mathbf{N}(x^n)/n$ differs from $W\mathbf{P}_X$ by $k\gamma$ in at least one component, where $\mathbf{N}(x^n), \mathbf{P}_X$ are column vectors consisting of $N(a|x^n), P_X(a)$ respectively for $a \in A$. Now if $y^n \in T_W^n(x^n,\gamma)$ then

$$\left|\frac{1}{n}N(b|y^n) - \sum_{a\in A}\frac{1}{n}N(a|x^n)W(b|a)\right| \leq |A|\gamma$$

for all $b \in B$. Thus we have that if $x^n \in D_X^n\left(\frac{\gamma}{k}\right)$ and $y^n \in T_W^n(x^n, \frac{\gamma}{3|A|})$ then $\mathbf{N}(y^n)/n$ differs from $\mathbf{P}_Y$ in some component by at least $\frac{2\gamma}{3}$. Thus $y^n \in D_Y^n(\frac{2\gamma}{3})$. We have shown that if $x^n \in D_X^n\left(\frac{\gamma}{k}\right)$ then $T_W^n(x^n, \frac{\gamma}{3|A|}) \subseteq D_Y^n(\frac{2\gamma}{3})$. But $P_{Y^n}\left(D_Y^n\left(\frac{2\gamma}{3}\right)\right) \to 0$ as $n \to \infty$ and $W^n\left(T_W^n(x^n, \frac{\gamma}{3|A|})|x^n\right) \to 1$ as $n \to \infty$ at a rate independent of x^n (From Lemma 2.12 in [5]. So if the input $\phi^n(Z^n)$ approximates the output due to X^n then we must have $P_{Z^n}\left(\phi^n(Z^n) \in D_X^n\left(\frac{\gamma}{k}\right)\right) \to 0$ as $n \to \infty$. Since if this were not so,

$$P_{\phi^n(Z^n)}W^n\left(D_Y^n\left(\frac{2\gamma}{3}\right)\right) - P_{Y^n}\left(D_Y^n\left(\frac{2\gamma}{3}\right)\right) > \beta$$

infinitely often, for some $\beta > 0$. This would imply that the input $\phi^n(Z^n)$ does not approximate the output due to X^n.

Now by a slight modification of Lemma 2.13 in [5]

$$\left|\frac{1}{n}\log\left|T_W^n\left(x^n, \frac{\gamma}{3|A|}\right)\right| - H(Y|X)\right| \leq \epsilon(\gamma)$$

for every $x^n \in D_{X^n}^c\left(\frac{\gamma}{k}\right)$ where $\epsilon(\gamma)$ is continuous in γ, independent of x^n and $\epsilon(\gamma) \to 0$ as $\gamma \to 0$. Define $Q^n = \left\{z^n : \frac{1}{n}h_{Z^n}(z^n) \leq c_n - \delta\right\}$. Clearly $|Q^n| \leq exp(n(c_n - \delta))$. Also define R^n as the image of Q^n under the mapping ϕ^n and

$$S^n = \bigcup_{x^n \in R^n \cap D_{X^n}^c(\frac{\gamma}{k})} T_W^n\left(x^n, \frac{\gamma}{3|A|}\right).$$

Then we have $|S^n| \leq \exp n(c_n - \delta)\exp n(H(Y|X) + \epsilon(\gamma))$ and $P_{\phi^n(Z^n)}W^n(S^n) \geq \frac{\alpha}{2}$ for sufficiently large $n \in I$. This is because

$$W^n\left(T_W^n(x^n, \frac{\gamma}{3|A|})|x^n\right) \to 1$$

as $n \to \infty$ at a rate independent of x^n,

$$P_{Z^n}\left[\phi^n(Z^n) \in D_{X^n}^c(\frac{\gamma}{k})\right] \to 1$$

as $n \to \infty$ and

$$P_{Z^n}\left[\phi^n(Z^n) \in R^n\right] > \alpha$$

for all $n \in I$. Using Equation (9) we can upper-bound $|S^n|$ as

$$|S^n| \le \exp n\left(H(Y) - \frac{9\delta}{10} + \epsilon(\gamma)\right) \le \exp n\left(H(Y) - \frac{\delta}{2}\right)$$

for sufficiently large $n \in I$, where the second inequality holds if we chose $\gamma > 0$ sufficiently small. Thus we have $P_{Y^n}(S^n) \to 0$ as $n \to \infty$ (By the strong source coding theorem for i.i.d sources). Thus for any deterministic mapping ϕ^n we can find a set $S^n \subseteq B^n$ such that $P_{\phi^n(Z^n)}W^n(S^n) \ge \frac{\alpha}{2}$ for sufficiently large $n \in I$ and $P_{Y^n}(S^n) \to 0$ as $n \to \infty$. So $\mathbf{Z}$ cannot be an approximating source for the FRDMC $\mathbf{W}$.

$\square$

We note that Theorem 6 implies that if $\underline{H}(\mathbf{Z}) < S$ then the source $\mathbf{Z}$ is not an approximating source for a FRDMC channel with resolvability S.

References

[1] S. Vembu, S. Verdú, and Y. Steinberg, " The Source-Channel Separation Theorem Revisited," *IEEE Trans. Information Theory*, vol. 41, pp. 44-54, Jan 1995.

[2] S. Vembu, S. Verdú, "Generating Random Bits from an Arbitrary source: Fundamental Limits," *IEEE Trans. Information Theory*, vol. 41, pp. 1322-1332, Sept. 1995.

[3] T.S. Han, S. Verdú, "Approximation Theory of Output Statistics," *IEEE Trans. Information Theory*, vol. 39, pp. 752-772, May 1993.

[4] T.S. Han, S. Verdú, "Spectrum invariancy under output approximation for full rank discrete memoryless channels," Probl. Peredach. Inform. (in Russian), no. 2, pp, 101-118, 1993.

[5] I. Csiszár, Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981

[6] A. Feinstein, "A new basic theorem in information theory," *IRE Trans. Information Theory*, vol. IT-4, pp. 2-22, 1954.

An Asymptotic Solution to the Support Region of a Minimum Mean Square-Error Gaussian Quantizer

Sangsin Na
Dept. of Electronics Engineering
Ajou University
Suwon, Korea
sangsin@eecs.umich.edu

David L. Neuhoff
Dept. of Electrical Engineering and Computer Science
College of Engineering
The University of Michigan
Ann Arbor, MI 48109-2122
neuhoff@eecs.umich.edu

Abstract

This paper derives an analytic expression for the support region of a minimum mean square-error Gaussian quantizer. The formula is obtained based on the Panter-Dite formula for the inner distortion and the exact expression of the outer distortion. The derived formula shows that the support region, defined by the two outmost quantization points, grows as $2\sqrt{6 \ln N}$, where N is the number of quantization points. This result contrasts that of the minimum mean square-error uniform quantizer, whose support grows as $2\sqrt{4 \ln N}$, revealing that the minimum mean-square error quantizer covers expectedly a wider input range than the uniform counterpart but the two support regions are linearly related. A corollary is that the overload distortion decreases with N^3.

Key words: asymptotic quantization theory, granular distortion, overload distortion, Gaussian source.

1 Introduction

An N-point scalar quantizer Q_N is characterized by quantization points $\{q_1, q_2, \dots, q_N\}$, quantization cells $\{S_1, S_2, \dots, S_N\}$ that form a partition of $\mathbb{R}$, and a quantization rule $Q_N(\cdot)$: $Q_N(\alpha) = q_i$ if $\alpha \in S_i$. The thresholds of Q_N are numbers $\{t_1, t_2, \dots, t_N, t_{N+1}\}$ with $t_1 = -\infty < t_2 < \cdots < t_N < t_{N+1} = \infty$. Then the quantization cells are given by $S_i = (t_i, t_{i+1}]$. The quantization error is the difference defined by $x - Q_N(x)$.

An N-point optimum scalar quantizer for a given source in the minimum mean-square error sense is a quantizer that minimizes the average distortion D

$$D = E\{(X - Q_N(X))^2\} = \sum_{i=1}^{N} \int_{S_i} (x - q_i)^2 p(x)\, dx = \sum_{i=1}^{N} \int_{t_i}^{t_{i+1}} (x - q_i)^2 p(x)\, dx,$$

where $p(x)$ is the probability density function of the source.

The distortion from the outmost quantization cells S_1 and S_N deserves special attention: these cells have unbounded quantization errors, i.e., $x - q_1$ and $x - q_N$ can be arbitrarily (positively or negatively) large for $x \in S_1$ or S_N. We will call this distortion the outer distortion and denote it by D_o. Similarly the distortion due to the rest of the cells, denoted D_i, will be called the inner distortion. (The notion of *overload distortion* is related with the outer distortion. Although there is no unanimous agreement, a portion of the outer distortion constitutes the overload distortion.)

Then D can be written as

$$D = \underbrace{\sum_{i=2}^{N-1} \int_{t_i}^{t_{i+1}} (x - q_i)^2 p(x)\, dx}_{D_i} + \underbrace{\int_{-\infty}^{t_2} (x - q_1)^2 p(x)\, dx + \int_{t_N}^{\infty} (x - q_N)^2 p(x)\, dx}_{D_o} .$$

We will pay special attention to the outmost quantization points q_1 and q_N due to their roles:

(a) They can be used as a good initial value to cut down the number of iterations in iterative and recursive design algorithms, particularly in a Lloyd-Max algorithm and derivatives thereof.

(b) They indicate the behavior of overload distortion. For example, they serve as an indicator of the amount of overload distortion. Therefore they can provide insight into the overall distortion and performance of the optimum quantizer.

In case of symmetric sources which result in optimum symmetric quantizers, the situation is further simplified since $q_N = -q_1$. Two most commonly-known sources, namely Gaussian and Laplacian sources, fall into this category [1]. For the roles that q_N plays, we will call it the *key parameter* of an optimum quantizer. (Although q_1 can play the same role, we have chosen q_N.) The support region of a quantizer will mean the interval defined by q_1 and q_N, i.e., $[q_1, q_N]$. The importance of the support region comes from the fact that the resulting quantization error is bounded or *granular* for the input falling in the region, i.e., the input values are well supported by the quantizer. The goal of this paper is to find a closed-form expression for the key parameter q_N and hence the support region of the optimum quantizer for a Gaussian source. The method presented herein can be applied to other sources such as a Laplacian source.

This paper derives an analytic expression for q_N (for large N) of an optimum Gaussian quantizer. The expression is derived from decomposing the total distortion into the inner and outer distortions, substituting the Panter-Dite formula [2] for the inner distortion and then minimizing the distortion over the choice of the threshold between the inner and outer distortion regions. This approach was first used by Algazi in [3] but was left to numerical results. A curve-fitting approach, different from Algazi's but based on the Panter-Dite formula, was used by Roe in [4]. This paper differs from Roe's in that the method is not of curve fitting and from Algazi's in that a closed-form analytic solution is presented herein. The resulting formula shows that q_N is asymptotically equal to $\sqrt{6 \ln N}$. The comparison with the true q_N reveals that the asymptotic formula has an increasing accuracy with N, which is of course expected: for $N \geq 32$ the formula stays within 1 % of the true values.

The contribution of the paper is that, for large N, (a) it provides an accurate analytic formula for q_N, which can be used for the design of an optimum Gaussian quantizer, and that (b) it enables the discussion of distortion without designing an optimum Gaussian quantizer. For example, it can show, when combined with the Panter-Dite formula, the proportions of granular and overload distortions and how the number of quantization points affects them. The formula is to be appreciated particularly because designing an optimum quantizer for large N can be avoided and, even when a quantizer is designed, it may not lend useful insight. Another possible application of the formula is in the discussion of quantizer mismatch, where an optimum quantizer is applied to a different source from the one that it is designed for.

2 Formulation and Asymptotic Solution

2.1 Formulation and Approach

For a symmetric N-point quantizer Q_N for the source with an even-symmetric density $p(x)$, the overall distortion D is written by

$$D = \sum_{i=2}^{N-1} \int_{t_i}^{t_{i+1}} \left(x - Q_N(x)\right)^2 p(x)\, dx + 2\int_{t_N}^{\infty} (x - q_N)^2 p(x)\, dx. \tag{1}$$

The optimality requires that the thresholds $t_2, \ldots, t_N$ and quantization points $y_1, \ldots, y_N$ should satisfy

$$\frac{\partial D}{\partial t_i} = 0 \text{ for } i = 2, \ldots, N, \text{ and } \frac{\partial D}{\partial q_i} = 0 \text{ for } i = 1, \ldots, N. \tag{2}$$

The parameter of our interest is t_N. Noting that $t_1 = -t_N$ due to the assumed symmetric quantizer, (1) is rewritten in terms of t_N as

$$D = \int_{-t_N}^{t_N} \left(x - Q_N(x)\right)^2 p(x)\, dx + 2\int_{t_N}^{\infty} (x - q_N)^2 p(x)\, dx. \tag{3}$$

And the optimality condition (2) implies that

$$\frac{\partial D}{\partial t_N} = 0. \tag{4}$$

The goal is to find an analytic solution to (4) using (3). However, there are a couple of difficulties in finding a solution directly from (3) and (4). First, solving (4) involves not only t_N, but also all the quantization points and the rest of the thresholds except the two extreme points t_1 and t_{N+1}. (So, there are $2N - 1$ other parameters to be considered. And symmetry reduces the number approximately to N.) Then finding a solution in this manner would be equivalent to essentially designing the quantizer, which one wants strongly to avoid. Secondly, an analytic solution is not probable due to mathematical complexity except for a very simple density, such as a uniform density.

Asymptotic Approach Due to the difficulties associated with the direct approach, we attempt the following approach, which we call the *asymptotic approach*:

(a) Express the two terms in (3) as the distortions of two optimum quantizers with $N - 2$- and 2-points, respectively, with the former for the inner distortion and the latter for the outer distortion.

(b) Then replace the inner distortion in (3) with the Panter-Dite approximation formula and the outer distortion with the explicit formula. This explicit formula is nothing but the functional relationship between the quantization point and the threshold of the 2-point optimum quantizer. (These replacements result in an expression for the distortion involving only one variable to be optimized over.)

For a positive u, let $D_i(u)$ and $D_o(u)$ be defined as follows:

$$D_i(u) = E\{(X - Q_{N-2}(X))^2 \,\big|\, |X| \le u\}P(|X| \le u) = \int_{-u}^{u} \left(x - Q_{N-2}(x)\right)^2 p(x)\,dx, \tag{5}$$

$$D_o(u) = E\{(X - Q_2(X))^2 \,\big|\, |X| > u\}P(|X| > u) = 2\int_{u}^{\infty} (x-q)^2 p(x)\,dx, \tag{6}$$

where $Q_{N-2}(\cdot)$ and $Q_2(\cdot)$ are, respectively, the minimum mean square-error quantizers with $N-2$ and 2 quantization points for the conditional distributions $p(x|C)$ and $p(x|C^c)$, where $C = \{x : -u \le x \le u\}$. In other words, $D_i(u)$ is the minimum distortion of any quantizer with $N-2$-points for the source input restricted to $[-u, u]$, and $D_o(u)$ with 2-points for the source input outside $[-u, u]$, weighted with appropriate probabilities.

Then the minimum distortion D^* of any N-point quantizer for $p(x)$ is related with $D_i(u)$ and $D_o(u)$ by the following expression:

$$D^* = \min_u D(u) = \min_u \left(D_i(u) + D_o(u)\right). \tag{7}$$

In the following we seek expressions for the terms in the objective function in (7). An approximation of $D_i(u)$ is given by the Panter-Dite formula.

$$\int_{-u}^{u} \left(x - Q_{N-2}(x)\right)^2 p(x)\,dx = \frac{1}{12N^2}\left(\int_{-u}^{u} p^{\frac{1}{3}}(x)\,dx\right)^3 \left(1 + o_u(1) + o_N(1)\right), \tag{8}$$

where $o_u(1)$ and $o_N(1)$ respectively stand for functions that vanish as $u \to \infty$ and $N \to \infty$. The expression for $D_o(u)$ is given by direct computation

$$D_o(u) = 2\int_{u}^{\infty} (x-q)^2 p(x)\,dx, \tag{9}$$

where $q = E\{X | X \ge u\}$.

Then (7) is expressed as follows:

$$D^* = \min_u \left\{ \frac{1}{12N^2}\left(\int_{-u}^{u} p^{\frac{1}{3}}(x)\,dx\right)^3 \left(1 + o_u(1) + o_N(1)\right) + 2\int_{u}^{\infty} (x-q)^2 p(x)\,dx \right\}. \tag{10}$$

The objective function in (10) is essentially the same as (13) in [3] with the main difference being the inclusion of $o_u(1)$ and $o_N(1)$ terms in (10).

The remainder of the paper is devoted to finding an approximate solution to (10) following the method described above in the case of a Gaussian source density.

2.2 Solution for a Gaussian Source

For a Gaussian density $p(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$, the two terms of the objective function in (7) or (10) are

$$D_i(u) = \frac{\sqrt{3}\pi}{2N^2}\left(1 - 2Q(\frac{u}{\sqrt{3}})\right)^3 \left(1 + o_u(1) + o_N(1)\right), \tag{11}$$

$$D_o(u) = 2\left(Q(u) - (q-u)p(u)\right). \tag{12}$$

Approximate Result One approximate and non-rigorous solution can be found by ignoring $o_u(1)$ and $o_N(1)$ terms and minimizing the resulting expression $\tilde{D}(u)$:

$$\tilde{D}(u) = \tilde{D}_i(u) + \tilde{D}_o(u),$$

where

$$\tilde{D}_i(u) \triangleq \frac{\sqrt{3}\pi}{2N^2}\Big(1 - 2Q(\frac{u}{\sqrt{3}})\Big)^3,$$
$$\tilde{D}_o(u) \triangleq 2\Big(Q(u) - (q-u)p(u)\Big).$$

We note that $\tilde{D}_o(u) = D_o(u)$. By straightforward differentiation

$$\frac{d\tilde{D}(u)}{du} = \frac{1}{\sqrt{2\pi}}\frac{3\pi}{N^2}\Big(1 - 2Q(\frac{u}{\sqrt{3}})\Big)^2 \exp(-\frac{u^2}{6}) - \frac{2}{\sqrt{2\pi}}(u-q)^2 \exp(-\frac{u^2}{2}). \tag{13}$$

Hence, setting (13) to 0 yields after a little bit of manipulation

$$\exp(\frac{u^2}{6})\Big(1 - 2Q(\frac{u}{\sqrt{3}})\Big) = \sqrt{\frac{2}{3\pi}}N(q-u). \tag{14}$$

The goal now is to find an analytic approximate solution to (14) for large N, or equivalently to

$$u^2 = 6\ln N + 3\ln\frac{2}{3\pi} + 6\ln(q-u) - 6\ln(1 - 2Q(\frac{u}{\sqrt{3}})), \tag{15}$$

which is obtained from (14) by taking the natural logarithm. Before solving (15) a couple of approximations are made for its last two terms in the right-hand side. First, from the asymptotic relationship between q and u in Appendices I and II, as $u \to \infty$,

$$\begin{aligned}\ln(q-u) &= \ln\frac{1}{u}(1 - \frac{2}{u^2} + \frac{10}{u^4} + O(\frac{1}{u^6}))\\ &= -\ln u + \ln\big(1 - \frac{2}{u^2} + \frac{10}{u^4} + O(\frac{1}{u^6})\big)\\ &= -\ln u + \big(-\frac{2}{u^2} + \frac{8}{u^4} + O(\frac{1}{u^6})\big).\end{aligned} \tag{16}$$

Secondly, also from Appendix II,

$$\ln(1 - 2Q(\frac{u}{\sqrt{3}})) = -\sqrt{\frac{6}{\pi}}\frac{1}{u}\exp(-\frac{u^2}{6}) + O(u^{-3}\exp(-\frac{u^2}{6})) \qquad (u \to \infty). \tag{17}$$

With these approximations, (15) becomes, as $u \to \infty$,

$$\begin{aligned}u^2 = {}& 6\ln N + 3\ln\frac{2}{3\pi} - 3\ln u^2 - 6(\frac{2}{u^2} - \frac{8}{u^4} + O(\frac{1}{u^6}))\\ &+ 6\sqrt{\frac{6}{\pi}}\frac{1}{u}\exp(-\frac{u^2}{6}) + O(u^{-3}\exp(-\frac{u^2}{6})),\end{aligned} \tag{18}$$

where the change $6\ln u = 3\ln u^2$ is made for convenience. Though one can solve (18) for an analytic asymptotic solution, a further approximation is helpful. A possible further

Table 1: The accuracy of the analytic solution

$R = \log_2 N$	True key parameter q_N	Solution to Eq. (14)	Analytic solution q in (21)	$\frac{q}{q_N}$
1	0.7979	1.0797	0.3890	0.4875
2	1.5104	1.6325	1.2971	0.8588
3	2.1520	2.1999	2.0463	0.9509
4	2.7326	2.7480	2.6797	0.9806
5	3.2607	3.2627	3.2330	0.9915
6	3.7441	3.7405	3.7283	0.9958
7	4.1897	4.1837	4.1796	0.9976
8	4.603535	4.5968	4.5962	0.9984
9	4.990612	4.9840	4.9846	0.9988
10	5.354963	5.3482	5.3497	0.9990
11	5.699835	5.6932	5.6949	0.9991
12	6.02783065	6.0216	6.0232	0.9992
13	6.34101629	6.3350	6.3366	0.9993
14	6.64091430	6.6354	6.6370	0.9994

approximation starts with noting the fact that, for any fixed (positive or negative) real ν, $\exp(-\frac{u^2}{6}) = o(u^\nu)$ as $u \to \infty$. This fact implies that the last two right-hand side terms in (18) are negligible compared with the terms involving $\frac{1}{u^2}$, $\frac{1}{u^4}$, and so on. Then (18) can be written as

$$u^2 = 6\ln N + 3\ln\frac{2}{3\pi} - 3\ln u^2 - \frac{12}{u^2} + \frac{48}{u^4} + O(\frac{1}{u^6}) \qquad (u \to \infty). \tag{19}$$

An analytic asymptotic solution is usually found in an iterative manner, as indicated in [6, pp. 11–14]: first one starts with a rough solution usually involving the most significant term and then makes refinements thereafter. Following this outlined approach, the asymptotic solution to (19) and the corresponding quantization point q are found and given by

$$u = \sqrt{6\ln N}\left(1 - \frac{\ln\ln N}{4\ln N} - \frac{\ln 9\pi}{4\ln N} - \frac{\ln^2\ln N}{32\ln^2 N} + \frac{2 - \ln 9\pi}{16}\frac{\ln\ln N}{\ln^2 N} + O(\frac{1}{\ln^2 N})\right) \tag{20}$$

$$q = \sqrt{6\ln N}\left(1 - \frac{\ln\ln N}{4\ln N} + \frac{2 - 3\ln 9\pi}{12\ln N} - \frac{\ln^2\ln N}{32\ln^2 N} + \frac{8 - 3\ln 9\pi}{48}\frac{\ln\ln N}{\ln^2 N} + O(\frac{1}{\ln^2 N})\right). \tag{21}$$

A. Accuracy of Solution To evaluate the accuracy of the formula, q from (21) is evaluated and compared with the numerical solution to (14) and the true key parameter q_N of the optimum Gaussian quantizer. The result is listed in Table 1 and plotted in Fig. 1, which shows that numerical results agree well.

First of all, the analytic and numerical solutions are close to each other and more so as R increases, as they should. More importantly, the ratio $\frac{q}{q_N}$ shows that the formula (21) stays within 1 % of true q_N for $R \geq 5$ and 0.1 % for $R \geq 10$. (For $R = 3$ and 4, it is within 5 and 2 %, respectively.) The tendency to smaller difference with increasing R is also observed, which implies that not only the ratio but the actual values get closer with the increasing number of quantization points. For these reasons, the analytic formula is very accurate and consequently q in (21) can be accepted practically as the true key parameter of the optimum quantizer for large R, for example $R \geq 5$.

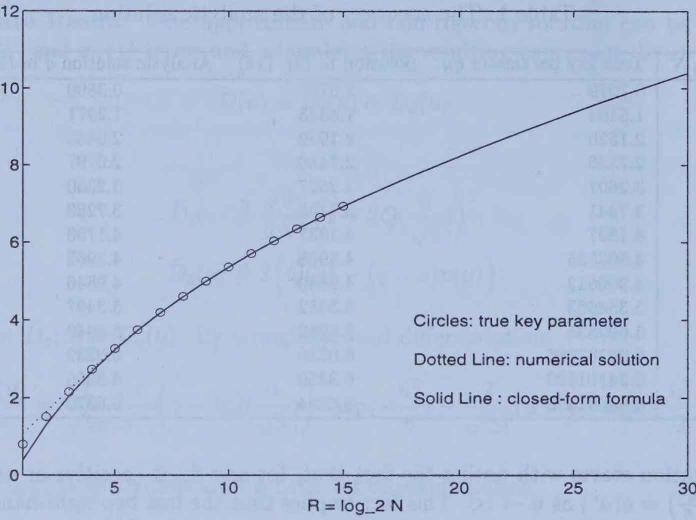


Figure 1: The accuracy of the analytic solution compared with the numerical solution and the actual quantizer: for $R \geq 5$ the analytic solution is within 1 % of the actual quantizer.

B. Comparison with Uniform Quantizer Accepting q in (21) practically as the true key parameter, it is informative to compare it with that of a uniform quantizer. Hui and Neuhoff [5] found through rigorous analysis that, for the minimum mean square-error uniform quantizer for a unit-variance Gaussian source, the outmost quantization point grows asymptotically with $\sqrt{4 \ln N}$. In contrast, (21) indicates that the optimum (nonuniform) quantizer grows asymptotically with $\sqrt{6 \ln N}$. This comparison shows that the optimum quantizer covers much wider a region than the uniform quantizer by $\sqrt{6 \ln N} - \sqrt{4 \ln N} = 0.4495\sqrt{\ln N}$ on each (positive and negative) side of the source, but that these support regions are linearly related, i.e., the ratio of the support regions of the optimum nonuniform to uniform quantizers $2\sqrt{6 \ln N}/2\sqrt{4 \ln N}$ converges to a constant $\sqrt{\frac{3}{2}}$. This will be a quantitative explanation that the overload distortion of a uniform quantizer is increasingly larger compared with that of an optimum quantizer, as the number of quantization points grows.

C. Formulas for Distortions The resulting $\tilde{D}_i^*$ and $\tilde{D}_o^*$ are now found from (11) and (12):

$$\tilde{D}_i^* = \frac{\sqrt{3}\pi}{2N^2}\left(1 - \frac{9}{N} + O\left(\frac{\ln \ln N}{N \ln N}\right)\right) \quad (N \to \infty), \tag{22}$$

$$\tilde{D}_o^* = \frac{3\sqrt{3}\pi}{2N^3}\left(1 - \frac{1}{6 \ln N} + O\left(\frac{\ln \ln N}{\ln^2 N}\right)\right) \quad (N \to \infty), \tag{23}$$

and

$$\frac{\tilde{D}_o^*}{\tilde{D}_i^*} = \frac{3}{N}\left(1 - \frac{1}{6\ln N} + O\left(\frac{\ln\ln N}{\ln^2 N}\right)\right) \qquad (N \to \infty). \tag{24}$$

These expressions show that the outer distortion $\tilde{D}_o^*$ is in the order of $\frac{1}{N^3}$ in contrast to the inner distortion $\tilde{D}_i^*$, which is in the order of $\frac{1}{N^2}$. It is noted that the explicit expression for $\tilde{D}_o^*$ has not been published in the literature. From these relationships, the approximate distortion $\tilde{D}^*$

$$\tilde{D}^* = \tilde{D}_i^* + \tilde{D}_o^* = \frac{\sqrt{3}\pi}{2N^2}\left(1 - \frac{6}{N} + O\left(\frac{\ln\ln N}{N\ln N}\right)\right) \qquad (N \to \infty). \tag{25}$$

Therefore, for large N, the optimum $\tilde{D}^* \approx \frac{\sqrt{3}\pi}{2N^2}$. Since the true optimum distortion $D^* \approx \tilde{D}^*$ for large N, we have the familiar formula for the distortion

$$D^* \approx \tilde{D}^* \approx \frac{\sqrt{3}\pi}{2N^2} \text{ for large } N, \tag{26}$$

which is also obtained (in fact originally) from the Panter-Dite formula [2].

3 Conclusion

This paper has derived an asymptotic formula for the key parameter based on the Panter-Dite formula, and hence for the support region of a minimum mean-square error quantizer, for the case of a Gaussian source. The derived formula is accurate for $N \geq 32$ and shows that the key parameter increases asymptotically with $\sqrt{6\ln N}$. This in turn reveals that an optimum nonuniform quantizer has a wider support region than the uniform counterpart but that their ratio converges to a constant. From this closed-form expression it is seen that the inner distortion decreases with N^2, as has been known for several decades, and that the outer distortion decreases with N^3. The formula can be useful in asymptotic study of issues involving quantization mismatch.

Appendix I

Consider a Gaussian random variable X with the density $p(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$. Then

$$q \triangleq E\{X|X \geq u\} = \frac{\displaystyle\int_u^\infty xp(x)\,dx}{\displaystyle\int_u^\infty p(x)\,dx} = \frac{\frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})}{Q(u)} = \frac{p(u)}{Q(u)}, \tag{27}$$

where $Q(u) = \int_u^\infty \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})\,dx$. Since

$$Q(u) \sim \frac{1}{\sqrt{2\pi}u}\exp(-\frac{u^2}{2})\sum_{n=0}^{\infty}(-1)^n\frac{(2n-1)!!}{u^{2n}} = \frac{p(u)}{u}\sum_{n=0}^{\infty}(-1)^n\frac{(2n-1)!!}{u^{2n}} \qquad (u \to \infty),$$

we have

$$q \sim \frac{u}{\sum\limits_{n=0}^{\infty}(-1)^n \frac{(2n-1)!!}{u^{2n}}} \qquad (u \to \infty), \tag{28}$$

where the symbol '$\sim$' denotes asymptotic equality, i.e., the ratio of the quantities of its sides approaches 1 as $u \to \infty$, and where $(2n-1)!! = 1 \cdot 3 \cdots (2n-1)$ and $(-1)!! = 1$ by definition. The first few terms of q are found through long division and given for future use:

$$q = u + \frac{1}{u} - \frac{2}{u^3} + \frac{10}{u^5} - \frac{74}{u^7} + O(\frac{1}{u^9}) \qquad (u \to \infty), \tag{29}$$

and

$$q - u = \frac{1}{u}(1 - \frac{2}{u^2} + \frac{10}{u^4} - \frac{74}{u^6} + O(\frac{1}{u^8})) \qquad (u \to \infty). \tag{30}$$

Therefore, for large u, $q \approx u + \frac{1}{u}$.

We note that, for a Laplacian source with $p(x) = \frac{1}{\sqrt{2}}\exp(-\sqrt{2}|x|)$,

$$q \triangleq E\{X | X \geq u\} = u + \frac{1}{\sqrt{2}}$$

for *any* nonnegative u. The equality is exact.

Appendix II

The Taylor series representation of $\ln(1+t)$ for $|t| < 1$ is given by

$$\ln(1+t) = t - \frac{t^2}{2} + \frac{t^3}{3} - + \cdots = \sum_{n=1}^{\infty}(-1)^{n+1}\frac{t^n}{n}.$$

Then, with the substitution $t = -\frac{2}{u^2} + \frac{10}{u^4} - \frac{74}{u^6} + O(\frac{1}{u^8})$, the following expression is obtained through multiplying out and then collecting terms:

$$\ln\big(1 - \frac{2}{u^2} + \frac{10}{u^4} - \frac{74}{u^6} + O(\frac{1}{u^8})\big) = -\frac{2}{u^2} + \frac{8}{u^4} - \frac{170}{3}\frac{1}{u^6} + O(\frac{1}{u^8}) \qquad (u \to \infty).$$

In the similar manner,

$$\ln\big(1 - 2Q(\tfrac{u}{\sqrt{3}})\big) = -2Q(\tfrac{u}{\sqrt{3}}) + O\left(Q^2(\tfrac{u}{\sqrt{3}})\right) \qquad (u \to \infty). \tag{31}$$

Since

$$Q(\tfrac{u}{\sqrt{3}}) \sim \sqrt{\frac{3}{2\pi}}\frac{1}{u}\exp(-\frac{u^2}{6})\sum_{n=0}^{\infty}\frac{(-3)^n(2n-1)!!}{u^{2n}} \tag{32}$$

$$= \sqrt{\frac{3}{2\pi}}\frac{1}{u}\exp(-\frac{u^2}{6})\big(1 + O(\frac{1}{u^2})\big) \qquad (u \to \infty). \tag{33}$$

The $Q^2(\frac{u}{\sqrt{3}})$ is in the order of $\frac{1}{u^2}\exp(-\frac{u^2}{3})$, is exponentially smaller than $Q(\frac{u}{\sqrt{3}})$ and therefore can be ignored. Taking the first term in (33) we have

$$\ln\big(1 - 2Q(\tfrac{u}{\sqrt{3}})\big) = -\sqrt{\frac{6}{\pi}}\frac{1}{u}\exp(-\frac{u^2}{6}) + O(u^{-3}\exp(-\frac{u^2}{6})) \qquad (u \to \infty). \tag{34}$$

References

[1] P.E. Fleischer, "Sufficient conditions for achieving minimum distortion in a quantizer," *Int. Conv. Rec.,* Part 1, pp. 104–111, 1964.

[2] P.F. Panter and W. Dite, "Quantization distortion in pulse count modulation with nonuniform spacing of levels," *Proc. IRE,* pp. 44–48, Jan. 1951.

[3] V.R. Algazi, "Useful approximations to optimum quantization," *IEEE Trans. Communication Technology,* COM-14, pp. 297–301, June 1966.

[4] G.M. Roe, "Quantizing for minimum distortion," *IEEE Trans. Information Theory,* IT-10, pp. 384–385, Oct. 1964.

[5] D. Hui and D.L. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization," paper submitted to IEEE Trans. on Information Theory.

[6] F.W.J. Olver, *Asymptotics and Special Functions,* Academic Press, 1974.

LIST OF AUTHORS

Agrawal, R. ... 693
Aji, S. ... 15
Ali, K.A. ... 600
Allen, B.M. ... 27
Alouini, M.-S. ... 146
Alpay, M.E. ... 651
Amato, F. ... 499
Anantharam, V. ... 10, 272
Arapostathis, A. ... 1
Ariola, M. ... 499
Atkin, G.E. ... 294
Barg, A. ... 359
Barton, R.J. ... 640
Bauby, C.E. ... 734
Bell, A.M. ... 262
Berny, A. ... 177
Birk, Y. ... 683
Black, C.K. ... 724
Bloch, N. ... 683
Bohacek, S. ... 79
Bortoff, S.A. ... 630
Brassard, G. ... 723
Brodersen, B. ... 765
Chang, A. ... 598
Cho, S.-I. ... 610
Choi, J. ... 292
Chong, E.K.P. ... 282, 296
Chung, K.C. ... 340
Chung, W. ... 252
Collins, O.M. ... 240
Costello, Jr., D.J. ... 240
Cruz, R.L. ... 693
da Silva e Silva, P.J. ... 450
Dabak, A.G. ... 441
Davey, M.C. ... 220
Divsalar, D. ... 201
Doi, H. ... 713
Dorato, P. ... 499
Duncan, T.E. ... 4
Dunham, J. ... 167
Elmasry, G.F. ... 49
Erstad, R. ... 372
Ertin, E. ... 772
Fain, E.A. ... 556
Famolari, D. ... 546
Fan, H.H. ... 782
Faris, D. ... 600
Farley, C.T. ... 749
Feder, M. ... 800
Fields, J.E. ... 566
Fitz, M.P. ... 391
Frank, C.D. ... 411
Gaborit, P. ... 566
Gelfand, S.B. ... 328
Georghiades, C.N. ... 343
Geraniotis, E. ... 156
Gerstel, O. ... 662
Goeckel, D.L. ... 126
Goldsmith, A. ... 516
Gollamudi, S. ... 790
Goodman, D. ... 546
Gorbet, R.B. ... 630
Grimm, J. ... 391
Guess, T. ... 318
Guha, S. ... 620
Ha, I.-J. ... 610
Hamkins, J. ... 29
Hartmann, C.R.P. ... 69
Heikkinen, T. ... 673
Helleseth, T. ... 593, 598
Herdtner, J.D. ... 296
Hochwald, B.M. ... 431
Honig, M.L. ... 298
Horn, G. ... 15
Hosoe, S. ... 479
Hosur, S. ... 441
How, J.P. ... 600
Huang, Y.-F. ... 790
Huffman, W.C. ... 566
Humes, Jr., C. ... 450
Huq, A.T. ... 343
Jin, H. ... 201
Johnson, Jr., C.R. ... 252, 401
Juedes, D.W. ... 703
Ke, N.-P. ... 489
Kim, S.-J. ... 610
Kim, Y.-H. ... 610
Kram, R. ... 757
Krogmeier, J.V. ... 391
Kucur, O. ... 294
Kulkarni, S. ... 810
Kulkarni, V. ... 79
Kumar, P.V. ... 598
Kuo, A.D. ... 734
Lee, E.B. ... 470
Lee, G.H. ... 292
Lee, H. ... 765
Leonard, D.A. ... 584
Li, L. ... 516
Li, X. ... 782
Liberty, S.R. ... 460
Litsyn, S. ... 353
Liu, J.-T. ... 328
Liu, X. ... 39, 136
Lou, X. ... 25
Luo, J.C. ... 470
Lutz, J.H. ... 703
Lynch, K.M. ... 724
MacKay, D.J.C. ... 220
Mambo, M. ... 713
Mandayam, N. ... 546
Marzetta, T.L. ... 431
Massey, P.C. ... 240
McEliece, R.J. ... 15, 201

Mitra, U. 772
Miyazaki, T. 479
Mohan, C.K. 69
Moustakides, G.V. 421
Na, S. 820
Nascimento, V.H. 242
Nauratanmull, V. 640
Neuhoff, D.L. 820
Nomay, I. 167
Nounou, N. 59
Öberg, M. 230
Okamoto, E. 713
Okino, C.M. 693
Packard, A. 600
Panayirci, E. 343
Paré, T. 600
Pérez, L.C. 25
Pham, K.D. 460
Pless, V. 566
Poor, H.V. 421
Pratt, G.A. 739
Pratt, J.E. 739
Rajan, R. 693
Ramaswami, R. 662
Rao, R.R. 526
Ratasuk, R. 298
Reddy, V.U. 340
Reznic, Z. 800
Ricker, S.L. 97
Rosenthal, J. 576
Rudie, K. 97
Safonov, M.G. 79
Sain, M.K. 460
Sasaki, G. 662
Sayed, A.H. 242
Sayeed, A.M. 536
Schmidl, T. 441
Schniter, P. 401
Sengupta, A. 107
Sethares, W.A. 262
Shavitt, Y. 89
Shepp, L. 191
Shi, W. 338
Shi, Y.Q. 49
Shor, M.H. 651
Siegel, P.H. 230, 366
Simon, M.K. 146
Siwamogsatham, S. 772
Smarandache, R. 576
Spielman, D.A. 211
Srinivasan, M. 29
Stark, W.E. 117
Su, H.-J. 156
Suri, D. 620
Suzuki, I. 620
Svaiter, B.F. 450
Takeshita, O.Y. 240
Teuscher, C. 765
Thomas, J. 308
Treichler, J. 252
Tseng, D.-F. 381
Ulukus, S. 506
Üstünel, A.S. 187
Varanasi, M.K. 318, 556
Venkatesan, S. 10
Verdú, S. 810
Visotsky, E. 411
Viswanath, P. 272
Visweswariah, K. 810
Wesel, R.D. 39, 136, 338
Wilson, S.T. 220
Wolf, J.K. 366
Worthen, A.P. 117
Xu, M. 15
Yates, R.D. 506
Yongzhe, X. 340
Ytrehus, Ø 372
Zakai, M. 187
Zamir, R. 800
Zeng, O.Z. 69
Zhang, J. 282
Zhang, N. 765
Zhou, S. 359
Zinoviev, V. 593
Zoltowski, M.D. 381
Zorzi, M. 526

UNIVERSITY OF ILLINOIS-URBANA
3 0112 033279792